Privacy-Aware Data Integration for Enhanced Quantile Inference under Heterogeneity

Leheng Cai [1 2] **Qirui Hu** [1 3 4] **Shuyuan Wu** [1 3]

Abstract

Quantile estimation and inference play essential roles in diverse scientific and industrial applications, and their accuracy can often be enhanced by integrating auxiliary data from multiple sites. However, developing efficient aggregation methods for quantile inference under potential privacy constraints, particularly with heterogeneous datasets, remains challenging. To address these issues, we propose a systematic framework for quantile estimation and inference under potential local differential privacy (LDP). The key idea is to construct weighted estimators by adaptively aggregating quantile estimates from target and source sites. The adaptive weights are determined by minimizing the asymptotic variance, incorporating an additional ℓ_2 penalty to account for parameter shift. A parallel stochastic gradient descent algorithm under LDP constraints is developed for weight estimation and valid inference. Additionally, we introduce a conservative weighted estimator to ensure robust inference across diverse heterogeneous scenarios. Rigorous theoretical analysis establishes the consistency, normality, and effectiveness of the proposed methods. Extensive numerical studies and real data application corroborate our theoretical findings.

1. Introduction

Motivation. Quantile estimation and inference play a central role in many scientific and industrial applications (Chernozhukov & Fernández-Val, 2011; Huang et al., 2017; Kallus et al., 2024; Deuber et al., 2024; Yadlowsky et al., 2025). Quantiles provide robust summaries of outcome distributions, especially in the presence of heavy tails or extreme events. For example, financial institutions routinely use quantile-based risk measures such as value-at-risk to quantify and manage investment risk (Chen, 2008; Barbaglia et al., 2023). Because such quantities directly inform risk management and operational decisions, improving the accuracy and efficiency of quantile inference has attracted substantial attention (Cai et al., 2025a).

A natural way to enhance quantile estimation is to leverage auxiliary datasets collected by multiple organizations or sites, thereby integrating information beyond a single target dataset (Wang et al., 2019; Cai et al., 2024b; Han et al., 2025). However, many real-world datasets contain sensitive personal or proprietary information and are subject to ethical and legal protections (Dwork et al., 2006; Cai et al., 2024a). In practice, privacy requirements often differ across sites: hospital networks, financial consortia, and federated platforms (e.g., smartphones or autonomous vehicles) operate under heterogeneous jurisdictional and organizational rules. As a result, some sites may release statistics that satisfy record-level local differential privacy (LDP), while others may only share unperturbed summaries or aggregate statistics (Konečný et al., 2016; Hard et al., 2018; Li et al., 2020; Nguyen et al., 2022). This raises a fundamental question:

> How can we effectively integrate auxiliary data sources to improve quantile estimation and inference, while respecting site-specific privacy constraints through mechanisms such as differential privacy?

Challenges. A growing literature studies data integration and federated inference; see Section 2 for a detailed review. Nonetheless, existing approaches face several obstacles in the setting of privacy-aware quantile inference. First, many integration methods construct weighted estimators that combine estimators from the target and auxiliary datasets, choosing the weights by minimizing criteria related to asymptotic variance (Li et al., 2022a; Cai et al., 2024b). In quantile

[1]The authors are listed in alphabetical order with equal contribution. [2]Department of Statistics and Data Science, Tsinghua University, Beijing, China [3]School of Statistics and Data Science, Shanghai University of Finance and Economics, Shanghai, China [4]Institute of Big Data Research, Shanghai University of Finance and Economics, Shanghai, China. Correspondence to: Qirui Hu <huqirui@mail.shufe.edu.cn>.

Proceedings of the 43rd International Conference on Machine Learning, Seoul, South Korea. PMLR 306, 2026. Copyright 2026 by the author(s).

problems under LDP, however, commonly used variance estimation tools such as classical sample variance or plug-in estimators (Zhu et al., 2021; Li et al., 2023; Huang et al., 2022; Gu & Chen, 2023; Han et al., 2025; Guo et al., 2025) are no longer directly applicable, because the locally privatized data substantially alter the underlying data distribution. Even without privacy constraints, most existing data integration methods focus on mean-type parameters or the minimization of smooth loss functions (Li et al., 2013; Lee et al., 2017; Chen & Xie, 2014; Li et al., 2022a; 2023; Zhu et al., 2024), whereas quantile loss functions are non-smooth and violate the regularity assumptions those methods rely on.

Second, current integration strategies typically impose restrictive assumptions on the relationship between target and auxiliary data sources. Many works assume that source and target parameters are exactly identical (Lee et al., 2017; Duan et al., 2020; Zhu et al., 2021; Wang & Shen, 2024), or that their differences are separated by a margin bounded away from zero (Huang et al., 2022; Li et al., 2022a; Cai et al., 2024c;a). Such assumptions can be unrealistic in heterogeneous multi-site environments and do not guarantee valid inference across the diverse scenarios encountered in practice.

Contributions. To address these challenges, we develop a systematic framework for improving quantile estimation and inference when both the target and auxiliary source datasets may be subject to LDP constraints.

- Our main idea is to construct weighted quantile estimators that adaptively combine information from the target and source sites. The weights are chosen to minimize the asymptotic variance of the resulting estimator, with an additional ℓ_2 penalty that regularizes parameter shifts between target and source sites. To implement this strategy under LDP, we design a parallel stochastic gradient descent (PSGD) algorithm that estimates the optimal weights while respecting local privacy constraints and enabling valid downstream inference. We further introduce a conservative weighted estimator that yields robust inference across a broad range of heterogeneous scenarios.
- Methodologically, our work provides a general and flexible framework for privacy-aware quantile estimation and inference via data integration. The proposed family of weighted estimators can substantially improve both estimation accuracy and inferential validity for the target quantile parameter under suitable conditions, and the framework accommodates diverse forms of heterogeneity across sites.
- Theoretically, we establish rigorous guarantees for the proposed procedures with the typical non-smooth quantile loss function: (i) we prove consistency of the variance estimator obtained from the PSGD algorithm, providing a sound basis for the weighted estimators and their associated inference; (ii) we derive consistency and asymptotic normality of the resulting weighted quantile estimators under heterogeneous multi-site settings; and (iii) we show that, under mild conditions and whenever auxiliary sites contain useful information, our methods achieve strictly better estimation and inference performance than using the target site alone.

The rest of the paper is organized as follows. Section 2 provides a brief literature review on data integration and LDP. Section 3.1 introduces the methodologies and complete algorithms. Theoretical results are presented in Section 4. Extensive numerical results and real data analysis are given in Section 5, and additional simulations are listed in Section A. All technical lemmas and proofs can be found in Sections B and C.

2. Related Work

Data integration. Recently, statistical data integration methods have attracted growing interest. Under the assumption of parameter homogeneity, existing studies primarily develop aggregation strategies that minimize appropriately defined asymptotic variance criteria of parameter estimators (Li et al., 2013; Chen & Xie, 2014; Wang et al., 2019; Zhu et al., 2021; Gu & Chen, 2023). When potential parameter shift exists, aggregation strategies in the existing literature typically also consider biases between parameters from auxiliary sources and the target parameter to mitigate adverse effects (Li et al., 2022a; 2023; Cai et al., 2024b;a;c; Han et al., 2025). To mitigate privacy concerns, classical methods aggregate summarized statistics (e.g., parameter estimates) rather than raw data (Chen et al., 2006; Lee et al., 2017; Duan et al., 2020; Guo et al., 2025; Bai et al., 2024), while recent approaches incorporate differential privacy constraints to achieve stronger privacy guarantees (Cai et al., 2024b;a;c; Li et al., 2024). In particular, statistical data integration methods have also been extensively developed for quantile problems. The existing literature mainly focuses on estimation (Hu et al., 2021; Jiang & Yu, 2021; Tan et al., 2022; Pillutla et al., 2024; Wang & Shen, 2024; Shi et al., 2025). Recently, a few studies have also addressed inference problems (Huang et al., 2022; Bai et al., 2024).

Local differential privacy (LDP). Differential privacy (DP) bounds how much a statistic can change if one record is modified, formalizing "plausible deniability" (Dwork et al., 2006). Variants such as Rényi DP, zCDP, and concentrated DP sharpen composition and enabled releases like the 2020 U.S. Census. DP's Achilles' heel is its reliance on a trusted curator; breaches, subpoenas, or misconfigurations can expose raw data (Narayanan & Shmatikov, 2008). Local DP (LDP) removes that trust by randomizing data at the source,

generalizing randomized response (Kasiviswanathan et al., 2011; Duchi et al., 2013; Liu et al., 2024). Pan-DP further shows only locally perturbed data withstand repeated intrusions, aligning pan-DP with LDP (Amin et al., 2020). LDP is now deployed in Chrome telemetry, Safari domain statistics, and Windows Defender reporting. These applications demonstrate that curator-free privacy can coexist with high-utility analytics, spurring research on utility-optimal protocols, adaptive privacy budgeting, and federated inference.

3. Methodology

3.1. Problem Description

We begin by introducing the model setup and notation. Due to page limitations, a complete list of notation is provided in Appendix B. Consider a total of N observations stored across a fixed set of $K+1$ sites, indexed by $\{0, 1, \ldots, K\}$. Denote the sample size at the k-th site by n_k, with $\sum_{k=0}^{K} n_k = N$, and assume that $n_k \asymp N/K$. At each site k, observations $\{X_{k,t}\}_{t=1}^{n_k} \subseteq \mathbb{R}$ are independently generated from an unknown distribution $\mathcal{P}_k$. The parameter of interest at each site is the quantile at a specified level $\tau \in (0,1)$. Specifically, define the check loss function as: $\ell(x,\theta) = (x-\theta)(\tau - \mathbf{1}(x \le \theta))$. Then, the quantile parameter at the k-th site satisfies that:

$$\theta_k = \arg\min_{\theta} \mathbb{E}_{x\sim\mathcal{P}_k}\{\ell(x,\theta)\}. \tag{1}$$

To estimate the parameter θ_k in practice, one typically minimizes the empirical counterpart of the objective function, which is given by $\widehat{\theta}_k = \arg\min_\theta \sum_{t=1}^{n_k} \ell(X_{k,t},\theta)$. Under regular conditions, it is assumed that $\widehat{\theta}_k$ admits the following asymptotic rule:

$$\sqrt{n_k}\left(\widehat{\theta}_k - \theta_k\right) \xrightarrow{d} \mathcal{N}\left(0, \tau(1-\tau)/f_k^2(\theta_k)\right), \tag{2}$$

where f_k denotes the probability density function associated with the distribution $\mathcal{P}_k$. Various algorithms are available for solving this empirical optimization problem, facilitating both estimation and statistical inference for θ_k. The classical and simplest method is based on order statistics (Van der Vaart, 2000). Specifically, the quantile parameter θ_k at site k can be directly estimated by taking the corresponding empirical quantile. Inference is typically conducted by plugging in a density estimator, such as a kernel density estimator, for the unknown probability density function evaluated at θ_k. While this method is straightforward and efficient, it directly utilizes raw data, limiting its applicability in sensitive scenarios. An alternative classical approach is Averaged Stochastic Gradient Descent (ASGD) (Polyak & Juditsky, 1992; Chen et al., 2023). Starting from an initial estimator $\widehat{\theta}_{k,0}$, ASGD iteratively updates the estimator at each site k as follows:

$$\widehat{\theta}_{k,t+1} = \widehat{\theta}_{k,t} - \eta_{k,t}\left\{\mathbf{1}\left(X_{k,t+1} \le \widehat{\theta}_{k,t}\right) - \tau\right\}, \tag{3}$$

where $0 \le \eta_{k,t} \le 1$ denotes the learning rate. The final estimator is computed as the average of all iterates as $\widehat{\theta}_k = n_k^{-1}\sum_{t=1}^{n_k}\widehat{\theta}_{k,t}$. Statistical inference can be conveniently implemented using self-normalized methods (Li et al., 2022b; Lee et al., 2022). Compared to the order-statistics-based method, the iterative nature of ASGD makes it naturally amenable to incorporating privacy-preserving mechanisms. Specifically, we consider here the ASGD algorithm under LDP constraints (Liu et al., 2023). Before introducing the detailed algorithm, we first provide formal definitions for DP and LDP.

Definition 3.1 (DP, see Dwork et al. (2006))**.** A randomized algorithm $\mathcal{A}$, taking a dataset consisting of individuals as its input, is (ϵ,δ)-differentially private if, for any pair of datasets S and S' that differ in the record of a single individual and any event E, the following inequality holds: $\mathbb{P}[\mathcal{A}(S) \in E] \le e^{\epsilon}\mathbb{P}[\mathcal{A}(S') \in E] + \delta$. When $\delta = 0$, $\mathcal{A}$ is called ϵ-differentially private (ϵ-DP).

Definition 3.2 (LDP, see Joseph et al. (2019))**.** An (ϵ,δ)-randomizer $R : X \to Y$ satisfies (ϵ,δ)-LDP if, for any event E and any input data point $X \ne X'$, $\mathbb{P}[R(X) \in E] \le e^{\epsilon}\mathbb{P}[R(X') \in E] + \delta$.

Next, inspired by (Liu et al., 2023), we modify the classical ASGD procedure in (3) by incorporating a local randomization step into the binary indicator function $\mathbf{1}(X_{k,t+1} \le \widehat{\theta}_{k,t})$. To be more precise, at the t-th iteration, we issue a query to the private data point $X_{k,t+1}$. In response, with probability r_k, we receive the true binary indicator, and with probability $1-r_k$, we receive a random variable $v \sim \text{Bernoulli}(0.5)$; see detailed Algorithm A.1 in Appendix. This local randomization is the classical randomized response mechanism under LDP (Warner, 1965; Kasiviswanathan et al., 2011; Duchi et al., 2013). Since the Algorithm A.1 enjoys $(\epsilon_k, 0)$-LDP with $\epsilon_k = \log\{(1+r_k)/(1-r_k)\}$, our Algorithms 1 and A.2 have $(\max_k \epsilon_k, 0)$-LDP guarantees as a direct consequence. Here, the response rate r_k controls the level of privacy protection, with smaller values corresponding to stronger privacy guarantees. When $r_k = 1$, the method reduces to the standard non-private case. Since the observed binary variable is now a randomized version of the original indicator, it is necessary to execute bias-correction to ensure an unbiased gradient estimate. Let $\widehat{\zeta}_{k,t}$ denote the perturbed binary variable observed at iteration t and site k. Under this LDP mechanism, the iterative updating formula in (3) becomes:

$$\begin{aligned}\widehat{\theta}_{k,t+1} = \widehat{\theta}_{k,t} - \eta_{k,t}\Bigg\{&\frac{1+r_k-2r_k\tau}{2}\widehat{\zeta}_{k,t} \\ &-\frac{1-r_k+2r_k\tau}{2}\left(1-\widehat{\zeta}_{k,t}\right)\Bigg\}.\end{aligned} \tag{4}$$

Similarly as in the classical ASGD method, the final estimator is obtained by averaging the iterates over n_k steps. Statistical inference can then be performed using self-normalization techniques adapted to this LDP setting, see Liu et al. (2023); Cai et al. (2025b).

The methods described above provide feasible algorithms for solving the optimization problem (1) at each site k, with possible LDP constraints. Specifically, the classical order-statistics approach can be applied directly if privacy protection is unnecessary, whereas the ASGD-based method should be employed in an LDP setting. However, the parameter estimation and inference at each site can potentially be improved further by appropriately aggregating data across multiple sites, especially if other sites contain useful and relevant information, such as sharing the same underlying quantile parameter. Without loss of generality, we treat site 0 as the target site and the remaining K sites as source sites.

3.2. Adaptive Weighted Estimator

A natural way to utilize information from multiple sites is to combine the estimators from the target and source sites into a weighted estimator. Specifically, for each site k, let $\widehat{\theta}_k$ denote the estimator of θ_k derived by one of the previously discussed methods, and let w_k be the corresponding weight satisfying $w_k \geq 0$ for all $0 \leq k \leq K$ and $\sum_{k=0}^{K} w_k = 1$. We then define the weighted estimator

$\widehat{\theta}(\mathbf{w}) = \sum_{k=0}^{K} w_k \widehat{\theta}_k$, where $\mathbf{w} = \{w_k\}_{k=0}^{K}$. Our goal is to determine weights $\{w_k\}$ that maximize the efficiency of the weighted estimator while controlling for the negative impact arising from heterogeneity in data distributions and parameters between the target and source sites. To this end, we introduce the following loss function with respect to the weights $\mathbf{w}$:

$$\mathcal{L}(\mathbf{w}) = \sum_{k=0}^{K} w_k^2 \sigma_k^2 / n_k + \lambda \sum_{k=0}^{K} w_k^2 b_k^2,$$

where σ_k^2/n_k is the asymptotic variance of the estimator $\widehat{\theta}_k$. We will rigorously prove that $\sigma_k^2 = \left\{4r_k^2 f_k^2(\theta_k)\right\}^{-1}\{1 - r_k^2(2\tau-1)^2\}$ in subsequent theoretical analysis. The bias term $b_k = \theta_k - \theta_0$ represents the parameter shift of the k-th source site relative to the target site 0. The tuning parameter $\lambda \geq 0$ controls the trade-off between variance and bias. In particular, setting $\lambda = 0$ yields classical inverse-variance weighting (Zhu et al., 2021; Shi et al., 2023), while setting $\lambda = 1$ approximately corresponds to minimizing the mean squared error of the estimator (Li et al., 2023). Minimizing $\mathcal{L}(\mathbf{w})$ with respect to $\mathbf{w}$ yields the oracle weights $\mathbf{w}^* = \arg\min_{\mathbf{w}} \mathcal{L}(\mathbf{w})$, which has the following explicit closed-form solution, for $0 \leq k \leq K$

$$w_k^* = \left\{\sum_{j=0}^{K}\left(\frac{\sigma_j^2}{n_j} + \lambda b_j^2\right)^{-1}\right\}^{-1}\left(\frac{\sigma_k^2}{n_k} + \lambda b_k^2\right)^{-1}. \quad (5)$$

Here, oracle weights refer to the ideal weights that minimize the estimator's asymptotic variance under the assumption that the true site-specific variances are known. The weights in (5) are adaptive in that they automatically adjust to each source site's characteristics: sites with higher noise (e.g., lower response rates r_k) or larger parameter shifts (biases b_k) receive smaller weights, while more stable and better-aligned sites receive larger weights. This allows the weighted estimator to prioritize sources that are most informative for the target site.

In practice, the oracle weights in (5) involve unknown quantities and must be estimated from data. A natural estimator for the bias term is $\widehat{b}_k = \widehat{\theta}_k - \widehat{\theta}_0$. For the variance term σ_k^2, the classical plug-in estimator is often used when an explicit form is available (Han et al., 2025). However, in our setting some sites operate under LDP constraints, and the plug-in method becomes infeasible because the raw data needed to estimate f_k cannot be accessed. To overcome this limitation, we propose a PSGD approach that estimates σ_k^2 automatically when raw data are unavailable. The key idea is to partition each site's local data into multiple subsets and run independent SGD procedures, referred to as chains, on these subsets, enabling direct variance estimation.

Specifically, at each site k, the data $\{X_{k,t}\}_{t=1}^{n_k}$ are randomly divided into M_k subsets, each forming an i.i.d. SGD chain. Let $\{X_{k,t}^{(m)}\}_{t=1}^{\lfloor n_k/M_k \rfloor}$ denote the data in chain m for $1 \leq m \leq M_k$. For each chain, PSGD initializes the estimator $\widehat{\theta}_{k,0}^{(m)} = \widehat{\theta}_{k,0}$ for $0 \leq k \leq K$ and updates it iteratively as follows:

$$\begin{aligned}\widehat{\theta}_{k,t+1}^{(m)} = \widehat{\theta}_{k,t}^{(m)} - \eta_{k,t}\Bigg\{&\frac{1+r_k-2r_k\tau}{2}\widehat{\zeta}_{k,t}^{(m)} \\ &-\frac{1-r_k+2r_k\tau}{2}\left(1-\widehat{\zeta}_{k,t}^{(m)}\right)\Bigg\},\end{aligned} \quad (6)$$

where $\widehat{\zeta}_{k,t}^{(m)}$ denotes the locally randomized version of the indicator $\mathbf{1}(X_{k,t+1}^{(m)} \leq \widehat{\theta}_{k,t}^{(m)})$. After completing the iterations within each chain, we compute the chain-specific estimator by $\widehat{\theta}_k^{(m)} = (\lfloor n_k/M_k \rfloor)^{-1}\sum_{t=1}^{\lfloor n_k/M_k \rfloor}\widehat{\theta}_{k,t}^{(m)}$. The final estimator at site k is then obtained by averaging across the M_k chain-specific estimators: $\widehat{\theta}_k = M_k^{-1}\sum_{m=1}^{M_k}\widehat{\theta}_k^{(m)}$. Note that these chain-specific estimators $\{\widehat{\theta}_k^{(m)}\}$ are independent of each other. This inspires the following variance estimator:

$$\widehat{\sigma}_k^2 = (M_k - 1)^{-1}\sum_{m=1}^{M_k}\lfloor n_k/M_k \rfloor\left(\widehat{\theta}_k^{(m)} - \widehat{\theta}_k\right)^2. \quad (7)$$

Next, we estimate the oracle weights defined in equation (5) by replacing the unknown parameters σ_k^2 and b_k with their estimators $\widehat{\sigma}_k^2$ and $\widehat{b}_k$, respectively. We denote these estimated weights as $\{\widehat{w}_k\}$. Subsequently, the resulting weighted estimator for the target parameter θ_0 and corresponding variance estimator are obtained as

$$\widehat{\theta}_{\text{est}} = \sum_{k=0}^{K} \widehat{w}_k \widehat{\theta}_k, \quad \widehat{\sigma}_{\text{est}}^2 = \sum_{k=0}^{K} \frac{N}{n_k} \widehat{w}_k^2 \widehat{\sigma}_k^2.$$

We will theoretically show that $\widehat{\sigma}_{\text{est}}^2/N$ is a consistent estimator of $\text{Var}(\widehat{\theta}_{\text{est}})$. Thus, statistical inference for $\widehat{\theta}_{\text{est}}$ can be readily conducted by constructing a $(1-\alpha)$-confidence interval:

$$[\widehat{\theta}_{\text{est}} - \mathcal{Z}_{\alpha/2}\widehat{\sigma}_{\text{est}}/\sqrt{N}, \widehat{\theta}_{\text{est}} + \mathcal{Z}_{\alpha/2}\widehat{\sigma}_{\text{est}}/\sqrt{N}],$$

where $\mathcal{Z}_{\alpha/2}$ is the upper $(\alpha/2)$-quantile of the standard normal distribution. The complete pseudo code for the data integration algorithm is described below in Algorithm 1.

Algorithm 1 Privacy-Aware Quantile Inference via Data Integration

Input: Learning rates $\{\eta_{k,t}\}$, sample sizes $\{n_k\}$, $N = \sum_k n_k$, number of chains $\{M_k\}$, target quantile τ, truthful response rates $\{r_k\}$, tuning parameter λ, and significance level α.
Output: $\widehat{\theta}_{\text{est}}$ and $[\widehat{\theta}_{\text{est}} - \mathcal{Z}_{\alpha/2}\widehat{\sigma}_{\text{est}}/\sqrt{N}, \widehat{\theta}_{\text{est}} + \mathcal{Z}_{\alpha/2}\widehat{\sigma}_{\text{est}}/\sqrt{N}]$.
Initialization: set $\widehat{\theta}_{k,0}^{(m)} \leftarrow 0$ for all k, m.

for $0 \le k \le K$ **do**
  for $1 \le m \le M_k$ **do**
    for $1 \le t \le \lfloor n_k/M_k \rfloor - 1$ **do**
      Obtain the locally randomizer $\widehat{\zeta}_{k,t}^{(m)} = \text{LRC}\left(\widehat{\theta}_{k,t}^{(m)}, r_k, X_{k,t+1}^{(m)}\right)$ using Algorithm A.1.
      Compute $\widehat{\theta}_{k,t}^{(m)}$ according to equation (6).
    end for
    Compute the chain-specific estimator by $\widehat{\theta}_k^{(m)} = (\lfloor n_k/M_k \rfloor)^{-1} \sum_{t=1}^{\lfloor n_k/M_k \rfloor} \widehat{\theta}_{k,t}^{(m)}$.
  end for
  Compute the final estimator and corresponding variance estimator at site k by $\widehat{\theta}_k = M_k^{-1} \sum_{m=1}^{M_k} \widehat{\theta}_k^{(m)}$, $\widehat{\sigma}_k^2 = (M_k - 1)^{-1} \sum_{m=1}^{M_k} \lfloor n_k/M_k \rfloor (\widehat{\theta}_k^{(m)} - \widehat{\theta}_k)^2$.
end for
Compute the weighted estimator and its variance estimator by $\widehat{\theta}_{\text{est}} = \sum_{k=0}^{K} \widehat{w}_k \widehat{\theta}_k$ and $\widehat{\sigma}_{\text{est}}^2 = \sum_{k=0}^{K} N \widehat{w}_k^2 \widehat{\sigma}_k^2 / n_k$, where $\widehat{w}_k = \left\{ \sum_{j=0}^{K} \left(\frac{\widehat{\sigma}_j^2}{n_j} + \lambda \widehat{b}_j^2 \right)^{-1} \right\}^{-1} \left(\frac{\widehat{\sigma}_k^2}{n_k} + \lambda \widehat{b}_k^2 \right)^{-1}$.

3.3. Conservative Weighted Estimator

Algorithm 1 provides an adaptive weighting scheme capable of automatic inference under possible LDP constraints. Its validity, however, relies on conditions governing the parameter shifts between target and source sites. In particular, the bias b_k must be either negligible or clearly distinguishable, that is, much smaller or much larger than $N^{-1/2}$. This requirement is weaker than most assumptions in prior work (Shi et al., 2023; Gu & Chen, 2023; Han et al., 2025), but it omits an important intermediate regime where b_k is exactly of order $N^{-1/2}$.

In this regime, even oracle weights cannot ensure valid inference. The intuition is straightforward. Consider the case with a single source site ($K = 1$). When $b_1 \ll N^{-1/2}$, the resulting bias is dominated by the estimator's variance and can be ignored. When $b_1 \gg N^{-1/2}$, the oracle assigns almost no weight to site 1, mitigating its impact. But when $b_1 \asymp N^{-1/2}$, the bias is of the same order as the standard errors of $\widehat{\theta}_0$ and $\widehat{\theta}_1$. In this setting, the oracle cannot down-weight site 1 sufficiently, leaving the final estimator with a bias comparable to its variance and invalidating inference. Our proposed estimator necessarily inherits this limitation. Figure 1 depicts the contrast between cases where the bias is negligible and where it competes with the variance.

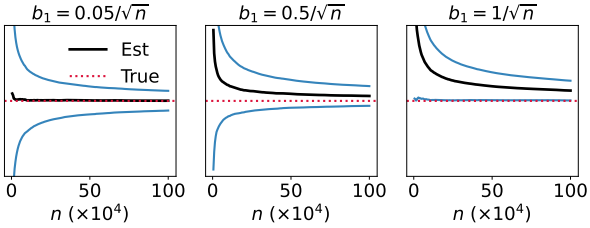


Figure 1. Illustration of confidence intervals under varying levels of bias. Each panel shows an estimator (black line) with bias b_1 and known variance $1/n$. The corresponding confidence interval with 95% significance level is represented by the blue lines. The red dashed line represents the true parameter value. The bias increases gradually from left to right. In the rightmost panel, the confidence interval no longer covers the true parameter, indicating that inference becomes invalid.

To address this issue, we propose a conservative approach based on the adaptive weighting method in Section 3.2. The key idea is to construct a conservative estimate of b_k, which tends to further down-weight biased source sites. Specifically, we replace the direct bias estimates with conservative upper bounds as

$$\widetilde{b}_k = \left|\widehat{b}_k\right| + \mathcal{C}\sqrt{\widehat{\sigma}_k^2/n_k + \widehat{\sigma}_0^2/n_0}, \quad \mathcal{C} > 0$$

for $k = 1, \ldots, K$, where $\widehat{\sigma}_k^2/n_k + \widehat{\sigma}_0^2/n_0$ is the asymptotic variance of $\widehat{b}_k$, and $\mathcal{C}$ controls the conservativeness level.

We then derive the conservative weights $\widetilde{w}_k$ for $0 \le k \le K$

by replacing σ_k^2 and b_k in equation (5) with $\widehat{\sigma}_k^2$ and $\widetilde{b}_k$, respectively, and setting $\widetilde{b}_0 = 0$.

Accordingly, the conservative weighted estimator for the target parameter θ_0 is defined as: $\widehat{\theta}_{\text{cons}} = \sum_{k=0}^{K} \widetilde{w}_k \widehat{\theta}_k$. The corresponding variance estimator for $\widehat{\theta}_{\text{cons}}$ is given by $\widehat{\sigma}_{\text{cons}}^2 = \sum_{k=0}^{K} N \widetilde{w}_k^2 \widehat{\sigma}_k^2 / n_k$. A complete algorithm is given in Algorithm A.2.

4. Theoretical Properties

In this section, we establish the theoretical properties for the proposed method. To this end, the following standard technical conditions are needed.

Assumption 4.1 (Property of f_k)**.** For every $0 \le k \le K$, $f_k(\cdot)$ is continuous and $f_k(\theta_k) > 0$. In addition, $|f_k'(\cdot)|$ is uniformly bounded by C for some constant $C > 0$.

Assumption 4.2 (Decaying learning rate)**.** The learning rate $\eta_{k,t}$ in equation (6) satisfies $\eta_{k,t} \asymp t^{-\beta}$ for some constant $\beta \in (1/2, 1)$.

Assumption 4.3 (Number of chains)**.** Assume that as $N \to \infty$, $M_k \to \infty$, and $M_k^{(\beta+1/2)\wedge(2-\beta)} \lesssim N^{(\beta-1/2)\wedge(1-\beta)}$.

Remark 4.4. Assumption 4.1 imposes standard regularity conditions on the probability density function. Assumption 4.2 requires a decaying learning rate, commonly assumed in the SGD literature (Polyak & Juditsky, 1992; Lee et al., 2022; Li et al., 2022b). Assumption 4.3 restricts the growth rate of the number of PSGD chains M_k relative to the sample size, ensuring accuracy of the final averaged estimator. Note that our main simulation designs are compatible with the distributional regularity assumptions.

We first provide the LDP guarantee for the PSGD algorithm, which is summarized in Proposition 4.5. Subsequently, the asymptotic properties of these estimators are established in Theorem 4.6.

Proposition 4.5 (Differential privacy)**.** *The PSGD algorithm for the k-th site satisfies $(\epsilon_k, 0)$-LDP with $\epsilon_k = \log\{(1+r_k)/(1-r_k)\}$. Therefore, both the Algorithms 1 and A.2 integrating $(K+1)$ sites LDP data are $(\max_{0\le k\le K} \epsilon_k, 0)$-LDP.*

Theorem 4.6 (PSGD algorithm)**.** *Under Assumptions 4.1 – 4.3, for each $k = 0, \ldots, K$, we have $\widehat{\sigma}_k^2 - \left\{4r_k^2 f_k^2(\theta_k)\right\}^{-1}\left\{1 - r_k^2(2\tau-1)^2\right\} = \mathcal{O}_p(1)$ and $\widehat{\theta}_k$ satisfies*

$$\sqrt{n_k}\left(\widehat{\theta}_k - \theta_k\right) \xrightarrow{d} \mathcal{N}\left(0, \frac{1 - r_k^2(2\tau-1)^2}{4r_k^2 f_k^2(\theta_k)}\right).$$

Theorem 4.6 extends the classical PSGD framework (Zhu et al., 2024) to the non-smooth quantile loss, delivers consistent variance estimation, and establishes the asymptotic normality of the estimator at the k-th site. We find that larger values of r_k lead to smaller asymptotic variance, but they also correspond to weaker privacy guarantees.

It is worth noting that our theory is not restricted to estimators produced by the PSGD algorithm; rather, it applies broadly to any procedure that provides consistent variance estimation. This requirement is summarized in Assumption 4.7.

Assumption 4.7 (Regularity of estimators)**.** For each $0 \le k \le K$, the estimators $\widehat{\sigma}_k^2$ and $\widehat{\theta}_k$ satisfy that $\widehat{\sigma}_k^2 - \sigma_k^2 = \mathcal{o}_p(1)$, and $\sqrt{n_k}\big(\widehat{\theta}_k - \theta_k\big) \xrightarrow{d} \mathcal{N}\big(0, \sigma_k^2\big)$ with $\sigma_k^2 = \left\{4r_k^2 f_k^2(\theta_k)\right\}^{-1}\left\{1 - r_k^2(2\tau-1)^2\right\}$.

Remark 4.8. Assumption 4.7 is a general regularity condition that accommodates various estimators, including for example, the proposed PSGD method under Assumptions 4.1 – 4.3, and the classical order statistics-based estimator in the non-private case (Van der Vaart, 2000).

Next, we provide theoretical guarantees for the weighted estimators. In our theoretical analysis, we investigate the properties of the proposed estimators under various choices of λ.

Specifically, under the assumptions stated below, Theorem 4.12 establishes the validity of statistical inference for these weighted estimators and efficiency improvement.

Assumption 4.9 (The choice of λ)**.** Assume that $\lambda = \mathcal{o}(1)$ and $\lambda|b_k| \gg N^{-1/2}$ for every $|b_k| \gg N^{-1/2}$.

Assumption 4.10 (Bias scale)**.** Assume the bias b_k for each $1 \le k \le K$ satisfies at least one of the following conditions:
(1) Vanishing bias: $|b_k| \ll N^{-1/2}$;
(2) Distinguishable bias: $|b_k| \gg N^{-1/2}$.

Remark 4.11. Assumption 4.9 specifies conditions on the tuning parameter λ. Assumption 4.10 allows the parameter shift to either vanish faster than $N^{-1/2}$ or be significantly larger than $N^{-1/2}$, covering a wide range of heterogeneous scenarios. This condition relaxes existing assumptions in the literature, which typically require either zero parameter shift (Zhu et al., 2021; Gu & Chen, 2023) or a parameter shift bounded away from zero (Li et al., 2022a; Han et al., 2025). Note that since the target parameter is the τ-quantile, the relevant source–target shift is $b_k = \theta_k - \theta_0$, measured relative to the statistical scale. Thus, local CDF closeness around the τ-quantile is sufficient for the bias-scale condition, while global distributional similarity is not required.

Theorem 4.12 (The property of $\widehat{\theta}_{\text{est}}$)**.** *Under Assumption 4.7, assume that one of the following three scenarios holds:*

(a) $\lambda = 0$, the bias scale satisfies Assumption 4.10 (1).

(b) λ is bounded away from 0, the bias scale satisfies Assumption 4.10 (2).

(c) λ satisfies Assumption 4.9, the bias scale satisfies Assumption 4.10.

Then we have (i) $\widehat{w}_k - w_k^* = \mathcal{O}_p(1)$ for $0 \le k \le K$, (ii)

$$\widehat{\sigma}^2_{\text{est}} - \sum_{k=0}^{K} (N/n_k) w_k^{*2} \frac{\{1 - r_k^2(2\tau-1)^2\}}{\{4r_k^2 f_k^2(\theta_k)\}} = \mathcal{O}_p(1),$$

(iii)

$$\left\{Avar(\widehat{\theta}_{est})\right\}^{-1/2} (\widehat{\theta}_{\text{est}} - \theta_0) \xrightarrow{d} \mathcal{N}(0,1),$$

$$\text{with } Avar(\widehat{\theta}_{est}) := \sum_{k=0}^{K} \frac{w_k^{*2}}{n_k} \frac{1 - r_k^2(2\tau-1)^2}{4r_k^2 f_k^2(\theta_k)}.$$

where w_k^* is the oracle weight defined in equation (5).

Further, under Assumptions 4.9, 4.10 and assume there exists $b_k \ll N^{-1/2}$ for some $1 \le k \le K$, the asymptotic variance of $\widehat{\theta}_{\text{est}}$ is strictly smaller than that of $\widehat{\theta}_0$. That is, as $N \to \infty$,

$$Avar(\widehat{\theta}_{est}) < Avar(\widehat{\theta}_0) := \frac{1 - r_0^2(2\tau-1)^2}{4r_0^2 f_0^2(\theta_0) n_0}.$$

Theorem 4.12 rigorously establishes the asymptotic normality of $\widehat{\theta}_{\text{est}}$ in the absence of unfavorable source sites whose biases are of order $N^{-1/2}$. By plugging the expression of the asymptotic variance into the weight w_k^*, we obtain

$$w_k^* \propto \left[\frac{1}{4n_k f_k^2(\theta_k)} \left\{ \frac{1}{r_k^2} - (2\tau-1)^2 \right\} + \lambda b_k^2 \right]^{-1},$$

which is shown that larger r_k or smaller b_k yield larger optimal weights w_k^* and hence a smaller asymptotic variance. Larger r_k, however, correspond to weaker privacy guarantees (Proposition 4.5), reflecting the fundamental accuracy–privacy trade-off inherent in the weighted estimator.

The asymptotic variance comparison in Theorem 4.12 further indicates that if at least one source site has a bias smaller than the order $N^{-1/2}$, the weighted estimator achieves a strictly smaller asymptotic variance than $\widehat{\theta}_0$. Hence, even if all the remaining r_j's ($j \ne k$) shrink to zero, the convergence rate of $\widehat{\theta}_{\text{est}}$ is then at least $\left(n_0 r_0^2 + n_k r_k^2\right)^{-1/2}$. As a result, the proposed method improves efficiency and yields narrower confidence intervals than relying solely on the target site data. Intuitively speaking, since the target-site estimator converges at rate $N^{-1/2}$, any source-site bias of smaller order becomes negligible relative to the asymptotic variance; incorporating information from such sites effectively enlarges the usable sample size and turns mild heterogeneity into a blessing.

In addition, it should be noted that our inference is constructed using normal quantiles, which yields narrower confidence intervals compared to self-normalization methods (Liu et al., 2023) at the same confidence level, even without data integration.

Finally, we summarize the asymptotic behaviour of the proposed conservative estimator $\widehat{\theta}_{\text{cons}}$.

Theorem 4.13 (The property of $\widehat{\theta}_{\text{cons}}$)**.** *Under Assumption 4.7, further assume that λ is bounded away from 0, $\mathcal{C} \to \infty$ and $\mathcal{C}\sqrt{|\widehat{\sigma}_k^2 - \sigma_k^2|} = \mathcal{O}_p(1)$ for $0 \le k \le K$ with $\sigma_k^2 = \left\{4r_k^2 f_k^2(\theta_k)\right\}^{-1}\left\{1 - r_k^2(2\tau-1)^2\right\}$, then we have $\sqrt{N/\widehat{\sigma}^2_{\text{cons}}}(\widehat{\theta}_{\text{cons}} - \theta_0) \xrightarrow{d} \mathcal{N}(0,1)$.*

Compared with Theorem 4.12, Theorem 4.13 imposes no additional bias-scale restrictions, allowing source sites whose biases are of order $N^{-1/2}$. A closer inspection of the estimator $\widehat{\theta}_{\text{cons}}$ shows that, regardless of the magnitude of b_k, the adjusted term $\widetilde{b}_k$ dominates $N^{-1/2}$ due to $C \to \infty$. This construction effectively prevents unfavorable source sites with bias of order $N^{-1/2}$ from affecting the transfer step, ensuring that the conservative method remains robust in more heterogeneous settings.

5. Experiments

5.1. Simulation setup

In this section, we examine the finite-sample performance of the proposed data integration method on both synthetic and real data. Code for implementing the proposed methods is publicly available[1]. In synthetic data, we fix quantile levels at $\tau \in \{0.25, 0.5, 0.75\}$. Data at each site are generated from either the Normal distribution $\mathcal{N}(\mu_k, 1)$ or the Cauchy distribution $\mathcal{C}(\mu_k, 1)$ with the target site fixed at $\mu_0 = 0$. We set the number of sites as $K = 3$ and the response rate to $r_k = 0.5$ for $0 \le k \le K$, yielding (1.10, 0)-LDP guarantees. The target sample size is $n_0 \in \{20000, 200000\}$, and each source sample size is three times larger. The number of local chains M_k varies between 8 and 20. The learning rate for chain m at site k in the t-th iteration is set as $\eta_{k,t_m} = 1/t_m^{0.6}$. Each experiment is replicated $R = 1,000$ times. We consider the following estimators for comparison:

ADP(1): The proposed adaptive weighted estimator with $\lambda = 1$.

ADP(decay): The proposed adaptive weighted estimator with $\lambda = 8 \log n_0 / \sqrt{n_0}$, which satisfies Assumption 4.9 when the shifts b_k are of constant order.

ADP(cons): The proposed conservative weighted estimator with $\lambda = 1$ and $\mathcal{C} = 1.96$.

[1] https://github.com/shuyuanwu/ldp-quantile-integration

IVW: The inverse variance weighted estimator, obtained as ADP with $\lambda = 0$, which ignores parameter shift.

Target (Liu et al., 2023): An ASGD-based estimator using only the target site's data under potential LDP constraints. Inference is conducted via self-normalization.

DPSGD (Song et al., 2013): Differentially private SGD, which adds noise directly to the stochastic gradients rather than using randomized response; Specifically, the update in (3) is modified to

$$\widehat{\theta}_{k,t+1} = \widehat{\theta}_{k,t} - \eta_{k,t}\left\{\mathbf{1}\left(X_{k,t+1} \le \widehat{\theta}_{k,t}\right) - \tau + Z_{k,t+1}\right\},$$

where $Z_t^k \sim \text{Laplace}(0, b)$ with scale $b = 1/\log\left\{(1 + r_k)/(1 - r_k)\right\}$. Here we set $\lambda = 1$.

Let $\theta^{(r)}$ denote the estimator in the r-th replication, and let $\text{CI}^{(r)}$ represent the corresponding 95% confidence interval (CI). We evaluate the estimators using (i) the mean squared error on the log scale (log MSE): $\log(R^{-1}\sum_{r=1}^{R}|\theta^{(r)} - \theta_0|^2)$; (ii) the empirical coverage probability at the 95% nominal level (ECP): $R^{-1}\sum_{r=1}^{R}\mathbf{1}\{\theta_0 \in \text{CI}^{(r)}\}$; and (iii) the averaged confidence interval length (length of CI): $R^{-1}\sum_{r=1}^{R}|\,\text{CI}^{(r)}\,|$. Complete implementation details can be found in Appendix A.1.

5.2. Simulation results

First, we evaluate finite-sample performance under scenarios of either vanishing or distinguishable bias, in order to verify Theorem 4.12. Specifically, for each $1 \le k \le 3$, we set μ_k to be either 0, $0.1/\sqrt{n_0}$, or $100/\sqrt{n_0}$. This setup creates 9 distinct bias levels ranging from complete homogeneity (level 1) to strong heterogeneity (level 9); see Table A.1 for detailed descriptions. Results are presented in Figures 2, A.1, A.2, and A.7(ii). We find that, first, with a fixed bias level, ADP methods yield decreasing log-MSE and CI width as n_0 grows, while ECP remains at or above 95%. In addition, ADP methods consistently outperform Target, with uniformly smaller log-MSE and narrower CIs. These empirical results align well with the theoretical guarantees in Theorem 4.12. Compared to competitors, IVW fails under large bias due to ignoring parameter shifts, and DPSGD produces larger MSE and wider CIs. These results illustrate the advantages of our approach.

Next, we examine the finite-sample performance under continuously varying bias values to verify Theorem 4.13. Specifically, for each $1 \le k \le 3$, we set $\mu_k = \mu$ and let μ vary from $\exp(-5)$ to 1. The results are summarized in Figures 3 and A.3. We observe that ADP(1) and ADP(decay) both deteriorate over a certain bias range, with MSE exceeding that of the Target estimator and ECP dropping far below 95%. In contrast, ADP(cons) remains uniformly robust. These findings demonstrate the usefulness of the conservative weighting strategy.

To further strengthen our simulation study, we conducted a set of sensitivity and robustness analyses, including (i) varying the SGD learning-rate hyper-parameter β; the response rates r_k; the number of sites K and the conservativeness level $\mathcal{C}$; (ii) using smaller numbers of chains M_0 and target sample size; (iii) considering more stringent privacy scenarios; (iv) an ADP(CV) rule for selecting the penalty parameter λ; (v) comparisons with a Lepski-type selector (Lepskii, 1991) and assumption-violation experiments; and (vi) wall-clock runtime comparisons. Detailed findings are provided in Appendix A.1.

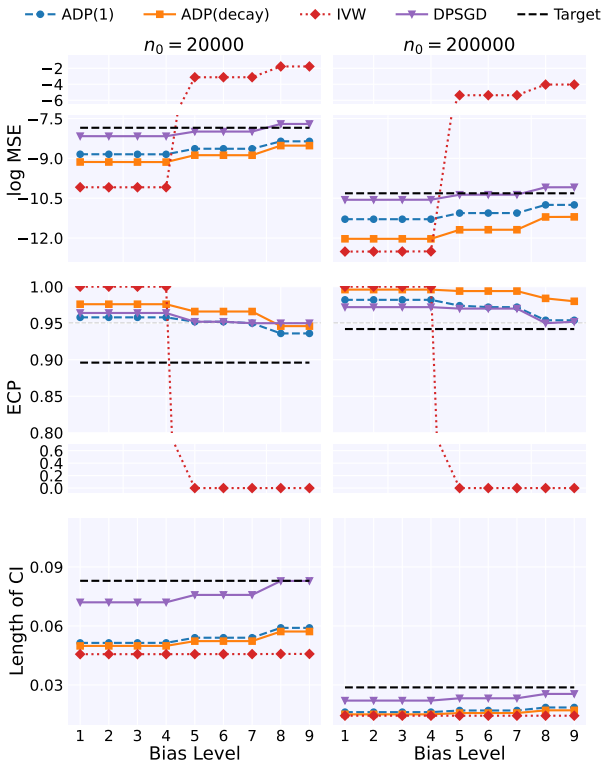


Figure 2. MSEs on the log scale (↓), empirical coverage probabilities at the 95% nominal level and the averaged confidence interval lengths (↓) under different heterogeneity scenarios. We consider scenarios of either vanishing or distinguishable bias. Bias levels range from complete homogeneity (level 1) to strong heterogeneity (level 9). The quantile level is fixed at $\tau = 0.5$. Data are generated from normal distributions, with the response rate fixed at $r_k = 0.5$ for all sites.

5.3. Real data

We evaluate our method on a real-world dataset widely employed in privacy research: the Government Salary Dataset (Plečko et al., 2024). This dataset is derived from the 2018 American Community Survey conducted by the U.S. Census Bureau and contains over 200,000 records with annual income (USD) as the response. Because income is sensitive personal financial information (Gillenwater et al., 2021), we treat it as privacy-protected data.

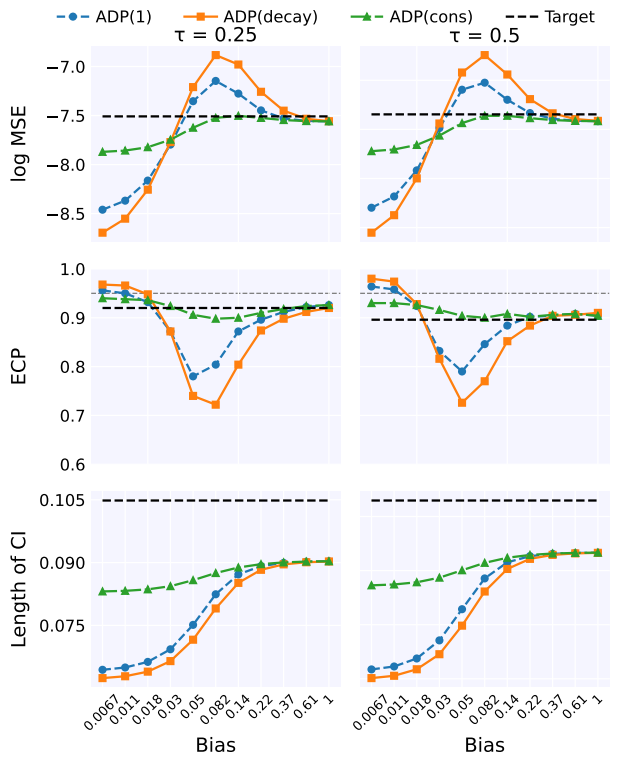


Figure 3. MSEs on the log scale (↓), empirical coverage probabilities at the 95% nominal level and the averaged confidence interval lengths (↓) under different heterogeneity scenarios. We consider continuously varying bias values from $\exp(-5)$ to 1. The target sample size is fixed at $n_0 = 20,000$. Data are generated from normal distributions, with the response rate fixed at $r_k = 0.5$ for all sites.

To reflect the dataset's geographic structure, we partitioned the data by "economic region", treating each region as a site. The original data include nine regions: we select one region (sample size 27,387) as the target site and use three larger regions (Southeast, Far West, and Mideast) as source sites. For all sites, we fix the number of chains at $M_k = 10$, all other hyperparameters follow Section 5. We log-transform the response for estimation and back-transform all reported quantities.

We target quantile levels $\tau \in \{0.25, 0.5\}$ and consider two privacy settings: (1) a homogeneous truthful response rate r for all sites, and (2) heterogeneous rates uniformly distributed on $[0.7, 1.0]$. We report point estimates (Est) and confidence-interval lengths (CI Len) for our two main approaches, ADP(decay) and ADP(cons), and include the Target estimator for reference. The results are summarized in Table 1. As expected, the Target estimator produces the longest intervals, while ADP(decay) yields the shortest in most cases. When privacy constraints are relaxed, the intervals for both ADP(decay) and ADP(cons) become shorter, consistent with our simulation findings. Across all scenarios, the confidence intervals of the Target estimator and those of the proposed ADP methods substantially overlap, indicating good practical performance on real data.

Table 1. Real-data analysis under different target quantile and truthful response rates. The number of chains per site is fixed at $M_k = 10$. $r_k = 0.7$ denotes a common truthful-response rate across sites, while "hetero r_k" denotes site-specific rates ranging from 0.7 to 1.0.

Metric	$\tau = 0.25$			$\tau = 0.5$		
	ADP (decay)	ADP (cons)	Target	ADP (decay)	ADP (cons)	Target
			$r_k = 0.7$			
Est	24909	23721	27442	44281	43403	45374
CI Len	1992	1910	3014	1041	1459	1950
			hetero r_k			
Est	24891	23724	27442	44268	44620	45374
CI Len	1753	1896	3014	1009	1410	1950

6. Concluding Remark

In summary, we propose a privacy-aware data-integration framework for quantile estimation and inference under local differential privacy. By combining PSGD-based variance estimation with adaptive bias–variance weighting, the proposed method improves efficiency when auxiliary sources are informative, while the conservative variant provides more robust inference under intermediate heterogeneity.

However, several limitations remain. First, our current asymptotic theory requires the number of PSGD chains M_k to diverge. This condition is used to establish consistency of the site-wise variance estimator via averaging across parallel chains; extending the theory to fixed-chain settings and developing non-asymptotic coverage guarantees are important directions for future work. Second, the present analysis assumes a fixed number of sites K, and extensions to diverging K remain open. Third, although the proposed framework is motivated by applications such as healthcare, our current experiments are primarily designed for methodological validation, and broader domain-specific evaluation in more complex real-world settings remains for future work.

Acknowledgments

The authors sincerely thank the anonymous reviewers, AC, and PCs for their valuable suggestions that have greatly improved the quality of our work. This work is supported by the Shanghai Engineering Research Center of Finance Intelligence (Grant No. 19DZ2254600). Leheng Cai's research is partially supported by China Association for Science and Technology and National Natural Science Foundation of China (No. 12171269). Shuyuan Wu's research is partially supported by National Natural Science Foundation of China (No. 12401392).

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none that we feel must be specifically highlighted here.

References

Amin, K., Joseph, M., and Mao, J. Pan-private uniformity testing. In Abernethy, J. and Agarwal, S. (eds.), *Proceedings of Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pp. 183–218. PMLR, 09–12 Jul 2020.

Bai, R., Zhang, Y., Yang, H., and Zhu, Z. Transfer learning for high-dimensional quantile regression with distribution shift. *arXiv preprint arXiv:2411.19933*, 2024.

Barbaglia, L., Consoli, S., and Manzan, S. Forecasting with economic news. *Journal of Business & Economic Statistics*, 41(3):708–719, 2023.

Cai, L., Hu, Q., Sun, J., and Wu, S. Time-uniform and asymptotic confidence sequence of quantile under local differential privacy. *Advances in Neural Information Processing Systems*, 38:114488–114520, 2025a.

Cai, L., Hu, Q., and Wu, S. Federated learning of quantile inference under local differential privacy. *arXiv preprint arXiv:2509.21800*, 2025b.

Cai, T., Chakraborty, A., and Vuursteen, L. Optimal federated learning for functional mean estimation under heterogeneous privacy constraints. *arXiv preprint arXiv:2412.18992*, 2024a.

Cai, T. T., Chakraborty, A., and Vuursteen, L. Federated nonparametric hypothesis testing with differential privacy constraints: Optimal rates and adaptive tests. *arXiv preprint arXiv:2406.06749*, 2024b.

Cai, T. T., Chakraborty, A., and Vuursteen, L. Optimal federated learning for nonparametric regression with heterogeneous distributed differential privacy constraints. *arXiv preprint arXiv:2406.06755*, 2024c.

Chen, L., Keilbar, G., and Wu, W. B. Recursive quantile estimation: Non-asymptotic confidence bounds. *Journal of Machine Learning Research*, 24(91):1–25, 2023.

Chen, S. X. Nonparametric estimation of expected shortfall. *Journal of financial econometrics*, 6(1):87–107, 2008.

Chen, X. and Xie, M.-g. A split-and-conquer approach for analysis of extraordinarily large data. *Statistica Sinica*, pp. 1655–1684, 2014.

Chen, Y., Dong, G., Han, J., Pei, J., Wah, B. W., and Wang, J. Regression cubes with lossless compression and aggregation. *IEEE Transactions on Knowledge and Data Engineering*, 18(12):1585–1599, 2006.

Chernozhukov, V. and Fernández-Val, I. Inference for extremal conditional quantile models, with an application to market and birthweight risks. *The Review of Economic Studies*, 78(2):559–589, 2011.

Csörgo, M. and Révész, P. *Strong approximations in probability and statistics*. Academic press, 1981.

Deuber, D., Li, J., Engelke, S., and Maathuis, M. H. Estimation and inference of extremal quantile treatment effects for heavy-tailed distributions. *Journal of the American Statistical Association*, 119(547):2206–2216, 2024.

Duan, R., Boland, M. R., Liu, Z., Liu, Y., Chang, H. H., Xu, H., Chu, H., Schmid, C. H., Forrest, C. B., Holmes, J. H., et al. Learning from electronic health records across multiple sites: A communication-efficient and privacy-preserving distributed algorithm. *Journal of the American Medical Informatics Association*, 27(3):376–385, 2020.

Duchi, J. C., Jordan, M. I., and Wainwright, M. J. Local privacy and statistical minimax rates. In *2013 IEEE 54th Annual Symposium on Foundations of Computer Science*, pp. 429–438. IEEE, 2013.

Dwork, C., Kenthapadi, K., McSherry, F., Mironov, I., and Naor, M. Our data, ourselves: Privacy via distributed noise generation. In *Annual International Conference on the Theory and Applications of Cryptographic Techniques*, pp. 486–503. Springer, 2006.

Gadat, S. and Panloup, F. Optimal non-asymptotic analysis of the ruppert–polyak averaging stochastic algorithm. *Stochastic Processes and their Applications*, 156:312–348, 2023.

Gillenwater, J., Joseph, M., and Kulesza, A. Differentially private quantiles. In Meila, M. and Zhang, T. (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 3713–3722. PMLR, 18–24 Jul 2021.

Gu, J. and Chen, S. X. Distributed statistical inference under heterogeneity. *Journal of Machine Learning Research*, 24(387):1–57, 2023.

Guo, Z., Li, X., Han, L., and Cai, T. Robust inference for federated meta-learning. *Journal of the American Statistical Association*, pp. 1–16, 2025.

Han, L., Hou, J., Cho, K., Duan, R., and Cai, T. Federated adaptive causal estimation (FACE) of target treatment

effects. *Journal of the American Statistical Association*, 120(551):1503–1516, 2025.

Hard, A., Rao, K., Mathews, R., Ramaswamy, S., Beaufays, F., Augenstein, S., Eichner, H., Kiddon, C., and Ramage, D. Federated learning for mobile keyboard prediction. *arXiv preprint arXiv:1811.03604*, 2018.

Hu, A., Jiao, Y., Liu, Y., Shi, Y., and Wu, Y. Distributed quantile regression for massive heterogeneous data. *Neurocomputing*, 448:249–262, 2021.

Huang, J., Wang, M., and Wu, Y. Estimation and inference for transfer learning with high-dimensional quantile regression. *arXiv preprint arXiv:2211.14578*, 2022.

Huang, Q., Zhang, H., Chen, J., and He, M. Quantile regression models and their applications: A review. *Journal of Biometrics & Biostatistics*, 8(3):1–6, 2017.

Jiang, R. and Yu, K. Smoothing quantile regression for a distributed system. *Neurocomputing*, 466:311–326, 2021.

Joseph, M., Mao, J., Neel, S., and Roth, A. The role of interactivity in local differential privacy. In *2019 IEEE 60th Annual Symposium on Foundations of Computer Science (FOCS)*, pp. 94–105, 2019. doi: 10.1109/FOCS.2019.00015.

Kallus, N., Mao, X., and Uehara, M. Localized debiased machine learning: Efficient inference on quantile treatment effects and beyond. *Journal of Machine Learning Research*, 25(16):1–59, 2024.

Kasiviswanathan, S. P., Lee, H. K., Nissim, K., Raskhodnikova, S., and Smith, A. What can we learn privately? *SIAM Journal on Computing*, 40(3):793–826, 2011.

Konečnỳ, J., McMahan, H. B., Ramage, D., and Richtárik, P. Federated optimization: Distributed machine learning for on-device intelligence. *arXiv preprint arXiv:1610.02527*, 2016.

Lee, J. D., Liu, Q., Sun, Y., and Taylor, J. E. Communication-efficient sparse regression. *Journal of Machine Learning Research*, 18(5):1–30, 2017.

Lee, S., Liao, Y., Seo, M. H., and Shin, Y. Fast and robust online inference with stochastic gradient descent via random scaling. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(7):7381–7389, 2022.

Lepskii, O. V. On a problem of adaptive estimation in gaussian white noise. *Theory of Probability & Its Applications*, 35(3):454–466, 1991. doi: 10.1137/1135065.

Li, M., Tian, Y., Feng, Y., and Yu, Y. Federated transfer learning with differential privacy. *arXiv preprint arXiv:2403.11343*, 2024.

Li, R., Lin, D. K., and Li, B. Statistical inference in massive data sets. *Applied Stochastic Models in Business and Industry*, 29(5):399–409, 2013.

Li, S., Cai, T. T., and Li, H. Transfer learning for high-dimensional linear regression: Prediction, estimation and minimax optimality. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 84(1):149–173, 2022a.

Li, T., Sahu, A. K., Talwalkar, A., and Smith, V. Federated learning: Challenges, methods, and future directions. *IEEE signal processing magazine*, 37(3):50–60, 2020.

Li, T., Shi, C., Wen, Q., Sui, Y., Qin, Y., Lai, C., and Zhu, H. Combining experimental and historical data for policy evaluation. In *Forty-first International Conference on Machine Learning*, 2023.

Li, X., Liang, J., Chang, X., and Zhang, Z. Statistical estimation and online inference via local sgd. In *Conference on Learning Theory*, pp. 1613–1661. PMLR, 2022b.

Liu, Y., Hu, Q., Ding, L., and Kong, L. Online local differential private quantile inference via self-normalization. In *International Conference on Machine Learning*, pp. 21698–21714. PMLR, 2023.

Liu, Y., Hu, Q., and Kong, L. Tuning-free estimation and inference of cumulative distribution function under local differential privacy. In *Forty-first International Conference on Machine Learning*, 2024.

Narayanan, A. and Shmatikov, V. Robust de-anonymization of large sparse datasets. In *2008 IEEE Symposium on Security and Privacy (sp 2008)*, pp. 111–125. IEEE, 2008.

Nguyen, A., Do, T., Tran, M., Nguyen, B. X., Duong, C., Phan, T., Tjiputra, E., and Tran, Q. D. Deep federated learning for autonomous driving. In *2022 IEEE Intelligent Vehicles Symposium (IV)*, pp. 1824–1830. IEEE, 2022.

Pillutla, K., Laguel, Y., Malick, J., and Harchaoui, Z. Federated learning with superquantile aggregation for heterogeneous data. *Machine Learning*, 2024. doi: 10.1007/s10994-023-06332-x.

Plečko, D., Bennett, N., and Meinshausen, N. fairadapt: Causal reasoning for fair data preprocessing. *Journal of Statistical Software*, 110:1–35, 2024.

Polyak, B. T. and Juditsky, A. B. Acceleration of stochastic approximation by averaging. *SIAM journal on control and optimization*, 30(4):838–855, 1992.

Serfling, R. J. *Approximation theorems of mathematical statistics*. John Wiley & Sons, New Jersey, 2009.

Shi, J., Wang, Y., Zhu, Z., and Lian, H. Decentralized learning of quantile regression: A smoothing approach. *Journal of Computational and Graphical Statistics*, pp. 1–11, 2025.

Shi, X., Pan, Z., and Miao, W. Data integration in causal inference. *Wiley Interdisciplinary Reviews: Computational Statistics*, 15(1):e1581, 2023.

Song, S., Chaudhuri, K., and Sarwate, A. D. Stochastic gradient descent with differentially private updates. In *2013 IEEE global conference on signal and information processing*, pp. 245–248. IEEE, 2013.

Tan, K. M., Battey, H., and Zhou, W.-X. Communication-constrained distributed quantile regression with optimal statistical guarantees. *Journal of Machine Learning Research*, 23(99):1–53, 2022. URL https://www.jmlr.org/papers/volume23/21-1223/21-1223.pdf.

Van der Vaart, A. W. *Asymptotic statistics*, volume 3. Cambridge university press, 2000.

Wang, C. and Shen, Z. Distributed high-dimensional quantile regression: estimation efficiency and support recovery. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 51415–51441, 2024.

Wang, X., Yang, Z., Chen, X., and Liu, W. Distributed inference for linear support vector machine. *Journal of machine learning research*, 20(113):1–41, 2019.

Warner, S. L. Randomized response: A survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association*, 60(309):63–69, 1965.

Xie, C., Jin, K., Liang, J., and Zhang, Z. Asymptotic time-uniform inference for parameters in averaged stochastic approximation. *arXiv preprint arXiv:2410.15057*, 2024.

Yadlowsky, S., Fleming, S., Shah, N., Brunskill, E., and Wager, S. Evaluating treatment prioritization rules via rank-weighted average treatment effects. *Journal of the American Statistical Association*, 120(549):38–51, 2025.

Zhu, W., Lou, Z., Wei, Z., and Wu, W. B. High confidence level inference is almost free using parallel stochastic optimization. *arXiv preprint arXiv:2401.09346*, 2024.

Zhu, X., Li, F., and Wang, H. Least-square approximation for a distributed system. *Journal of Computational and Graphical Statistics*, 30(4):1004–1018, 2021.

A. Additional Discussion and Results

Algorithm A.1 Locally Randomized Compare (Liu et al., 2023)

Require: Inquiry θ, response rate r, private data x
Ensure: A randomized binary response
1: Sample $u \sim \text{Bernoulli}(r)$
2: Sample $v \sim \text{Bernoulli}(0.5)$
3: **if** $u = 1$ **then**
4: **return** $\mathbf{1}\{x \leq \theta\}$
5: **else**
6: **return** v
7: **end if**

Algorithm A.2 Conservative Weighting Estimator

Input: Learning rates $\{\eta_{k,t}\}$, sample sizes $\{n_k\}$, number of chains $\{M_k\}$, target quantile τ, truthful response rates $\{r_k\}$, tuning parameter λ, and significance level α.
Output: $\widehat{\theta}_{\text{cons}}$ and $[\widehat{\theta}_{\text{cons}} - \mathcal{Z}_{\alpha/2}\widehat{\sigma}_{\text{cons}}/\sqrt{N}, \widehat{\theta}_{\text{cons}} + \mathcal{Z}_{\alpha/2}\widehat{\sigma}_{\text{cons}}/\sqrt{N}]$.
Initialization: set $\widehat{\theta}_{k,0}^{(m)} \leftarrow 0$ for all k, m.
for $0 \leq k \leq K$ **do**
for $1 \leq m \leq M_k$ **do**
for $1 \leq t \leq \lfloor n_k/M_k \rfloor - 1$ **do**
Obtain the locally randomizer $\widehat{\zeta}_{k,t}^{(m)} = \text{LRC}\left(\widehat{\theta}_{k,t}^{(m)}, r_k, X_{k,t+1}^{(m)}\right)$ using Algorithm A.1.
Compute $\widehat{\theta}_{k,t}^{(m)}$ according to equation (6).
end for
Compute the chain-specific estimator by $\widehat{\theta}_k^{(m)} = (\lfloor n_k/M_k \rfloor)^{-1} \sum_{t=1}^{\lfloor n_k/M_k \rfloor} \widehat{\theta}_{k,t}^{(m)}$.
end for
Compute the final estimator and corresponding variance estimator at site k by
$\widehat{\theta}_k = M_k^{-1} \sum_{m=1}^{M_k} \widehat{\theta}_k^{(m)}, \quad \widehat{\sigma}_k^2 = \frac{1}{M_k - 1} \sum_{m=1}^{M_k} \lfloor n_k/M_k \rfloor (\widehat{\theta}_k^{(m)} - \widehat{\theta}_k)^2$.
end for
Compute the conservative weighted estimator and its variance estimator by $\widehat{\theta}_{\text{cons}} = \sum_{k=0}^{K} \widetilde{w}_k \widehat{\theta}_k$ and $\widehat{\sigma}_{\text{cons}}^2 = \sum_{k=0}^{K} N \widetilde{w}_k^2 \widehat{\sigma}_k^2 / n_k$, where $\widetilde{w}_k = \left\{ \sum_{j=0}^{K} \left(\frac{\widehat{\sigma}_j^2}{n_j} + \lambda \widetilde{b}_j^2 \right)^{-1} \right\}^{-1} \left(\frac{\widehat{\sigma}_k^2}{n_k} + \lambda \widetilde{b}_k^2 \right)^{-1}$ and $N = \sum_k n_k$.

A.1. Other results in section 5 and complete experimental details.

Experimental details. We provide here additional experimental details not described in the main text. For all experiments, we use 36 Intel(R) Xeon(R) Gold 6271 CPUs, equipped with a total of 128GB of RAM and 500GB of storage. The experiments are implemented using Python 3.12, and the computational time required to generate each figure is approximately 3 to 8 hours.

A.2. Sensitivity and robustness analysis

Unless otherwise specified in this subsection, the target sample size is fixed at $n_0 = 20{,}000$, and the data are generated from normal distributions with a fixed response rate $r_k = 0.5$ for all sites. All other settings follow Section 5.1.

Sensitivity analysis. First, we investigate the effect of the learning-rate schedule. Table A.2 summarizes the ECPs. We find that the ECPs remain comparable across different values of β, indicating that the proposed estimator is stable with respect to the choice of the learning-rate schedule.

We also study the penalty parameter λ and the conservative parameter $\mathcal{C}$. Besides the default decay rule $\lambda = 8\log(n_0)/\sqrt{n_0}$, we implement ADP(CV), a privacy-free repeated two-fold cross-validation rule based only on existing chain-level summaries. Specifically, the chains are randomly split into two equal halves; adaptive weights are fitted on the training

Table A.1. Bias levels for source sites. Target site has $b_0 = 0$.

Level	Description of source sites	Biases $b_k,\ k = 1, \ldots, 3$
1	All source sites unbiased	$b_1 = b_2 = b_3 = 0$
2	One weakly biased source	$b_1 = b_2 = 0,\ b_3 = 0.1/\sqrt{n_0}$
3	Two weakly biased sources	$b_1 = 0,\ b_2 = b_3 = 0.1/\sqrt{n_0}$
4	Three weakly biased sources	$b_1 = b_2 = b_3 = 0.1/\sqrt{n_0}$
5	One strongly biased source	$b_1 = b_2 = 0,\ b_3 = 100/\sqrt{n_0}$
6	One weak + one strong	$b_1 = 0,\ b_2 = 0.1/\sqrt{n_0},\ b_3 = 100/\sqrt{n_0}$
7	Two weak + one strong	$b_1 = b_2 = 0.1/\sqrt{n_0},\ b_3 = 100/\sqrt{n_0}$
8	Two strongly biased sources	$b_1 = 0,\ b_2 = b_3 = 100/\sqrt{n_0}$
9	One weak + two strong	$b_1 = 0.1/\sqrt{n_0},\ b_2 = b_3 = 100/\sqrt{n_0}$

half using $\lambda(c; n_{\mathrm{tr}}) = c \log(n_{\mathrm{tr}})/\sqrt{n_{\mathrm{tr}}}$, evaluated on the validation half using a debiased held-out AMSE criterion, repeated over 10 random half-and-half splits, and then refitted on the full data using the selected c. We search over $c \in \{0.25, 0.5, 1, 2, 4, 8, 16\}$. Because this procedure is pure post-processing of already released chain-level summaries, it incurs no additional privacy budget. Figure A.8(i) shows that ADP(CV) performs comparably to ADP(decay). For $\mathcal{C}$, our default is 1.96, the two-sided 95% z-critical value; the sensitivity analysis over $\mathcal{C} \in \{1.64, 1.96, 2.33, 2.58\}$ is broadly comparable across choices, as shown in Figure A.8(ii).

We further compare ADP(cons) with a simple Lepski-type hard-selection baseline (Lepskii, 1991). The two methods are related in spirit because both use variance information to control the bias–variance tradeoff, but the mechanisms differ: a classical Lepski rule selects one candidate from an ordered family after threshold calibration, whereas ADP(cons) inflates each estimated source shift by a variance-dependent term and then continuously downweights heterogeneous sources. Figure A.9 shows that, in the moderate-bias regime, the Lepski-type baseline is more aggressive and exhibits clear undercoverage, while ADP(cons) is more stable.

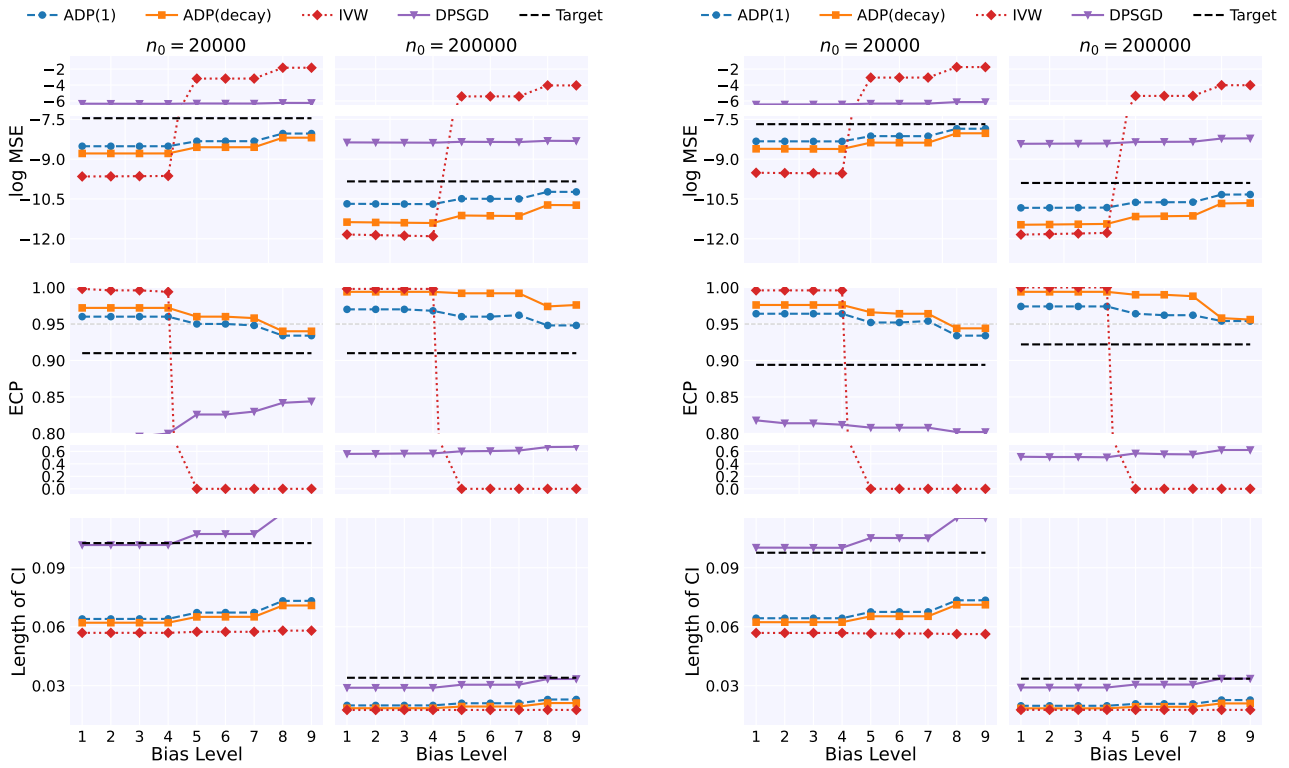
Next, we conduct a sensitivity analysis on the truthful response rate. We focus here on a homogeneous setting by fixing $\mu_k = 0$ for all $0 \le k \le 3$. We consider three target response rates, $r_0 \in \{1, 0.4, 0.25\}$, corresponding respectively to no privacy, moderate local privacy, and strong local privacy at the target site. For each target scenario, we vary the source-site response rates from 0.25 to 0.9, while keeping all other settings identical to Section 5.1. The results are summarized in Figure A.4.

We observe the following patterns. First, in almost all settings, the ADP estimators consistently improve upon the Target baseline, yielding smaller MAE, narrower confidence intervals, and empirical coverage close to (or above) the nominal 95% level. The improvement becomes more pronounced as the source response rate increases. In addition, when the target site is non-private ($r_0 = 1$), the benefit of integrating source sites is quite limited when their response rates are relatively small.

Robustness analysis. First, we vary the number of sites K, thereby examining the estimator's performance when incorporating information from fewer or more auxiliary sources. Second, we evaluate the impact of using smaller numbers of local chains M_0 to assess sensitivity with respect to variance estimation. Third, we investigate more stringent local-privacy scenarios by setting $r_k = 0.25$ for all sites, corresponding to strong privacy protection across the entire system. Detailed results are provided in Figures A.5, A.6, and A.7(i). Overall, the findings are qualitatively consistent with those in the main text, further demonstrating the robustness of the proposed method.

Figure A.9(ii) further considers $n_0 \in \{5{,}000, 10{,}000\}$, with qualitative conclusions remaining the same as in the main text. For privacy strength, the main setting is $r_k = 0.5$, while Figure A.4 studies target response rates $r_0 \in \{1, 0.4, 0.25\}$ and source response rates from 0.25 to 0.9; Figures A.6 and A.7(i) study the stronger-privacy case $r_k = 0.25$ for all sites. For chain number, the main baseline uses $M_k \in [8, 20]$, and Figure A.5(iii) considers $M_0 = 6$. These settings support reasonably stable empirical calibration.

We also add two robustness settings that intentionally violate Assumption 4.1, while keeping the rest of the setup the same as in Section 5.1. The first uses a Laplace design, replacing the error distribution with $\varepsilon_{k,t} \sim \mathrm{Laplace}(0, 1)$; the density has a cusp at the median, so the smoothness requirement in Assumption 4.1 is violated. Figure A.10(i) shows that the qualitative conclusions are broadly consistent with those in the main vanishing/distinguishable-bias experiment. The second uses a

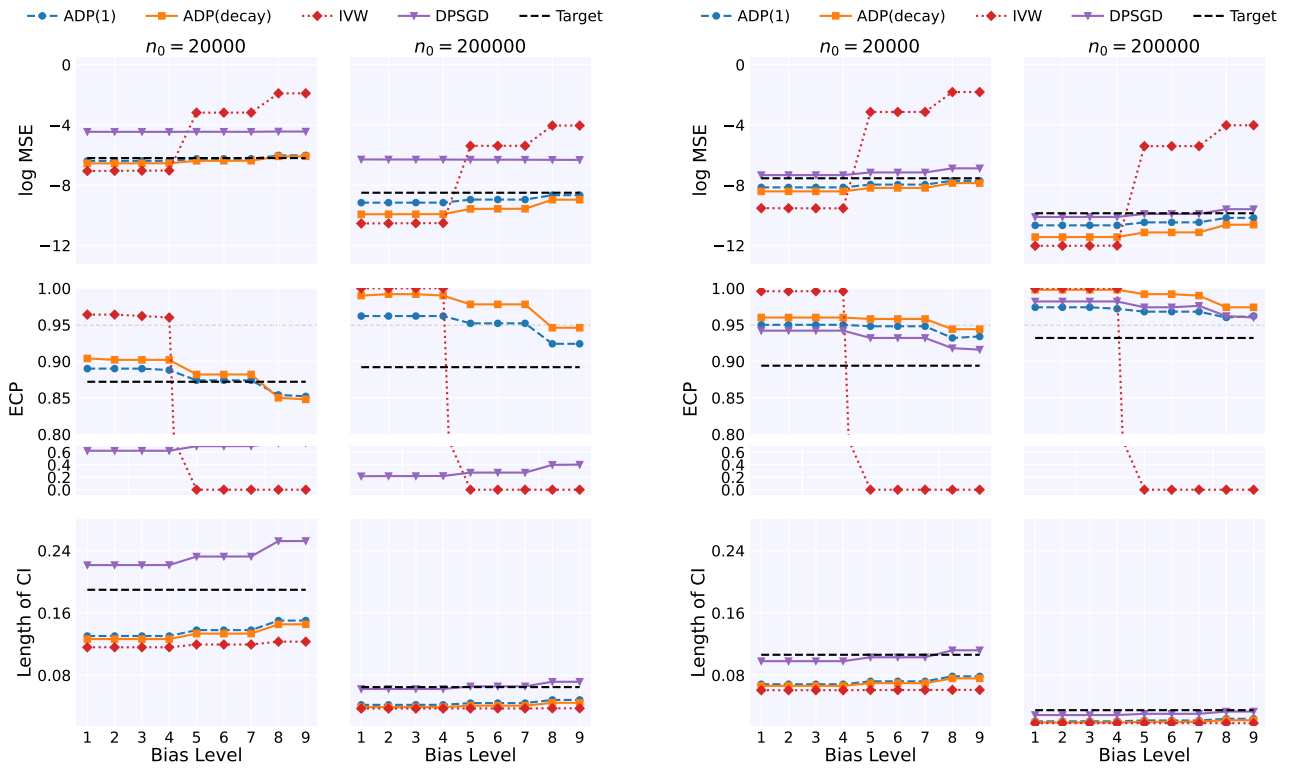


(i) $\tau = 0.25$ (ii) $\tau = 0.75$

Figure A.1. MSEs on the log scale (↓), empirical coverage probabilities at the 95% nominal level and the averaged confidence interval lengths (↓) under scenarios of either vanishing or distinguishable bias. Bias levels range from complete homogeneity (level 1) to strong heterogeneity (level 9). Data for both panels are generated from normal distributions, with the response rate fixed at $r_k = 0.5$ for all sites.

censored-normal design, where latent normal observations are censored at zero; this introduces a point mass at the target quantile and violates the continuity/absolute-continuity requirement in Assumption 4.1. Figure A.10(ii) shows that, in this case, empirical coverage deteriorates substantially.

Finally, we report a time-cost comparison under a representative setting matching the main simulation in Section 5.1. Figure A.11 gives serial wall-clock time per replication for each method. ADP(1), ADP(decay), and IVW have nearly identical runtimes, DPSGD is substantially slower, and Target is the fastest.



(i) $\tau = 0.25$ (ii) $\tau = 0.5$

Figure A.2. MSEs on the log scale (↓), empirical coverage probabilities at the 95% nominal level and the averaged confidence interval lengths (↓) under scenarios of either vanishing or distinguishable bias. Bias levels range from complete homogeneity (level 1) to strong heterogeneity (level 9). Data for both panels are generated from Cauchy distributions, with the response rate fixed at $r_k = 0.5$ for all sites.

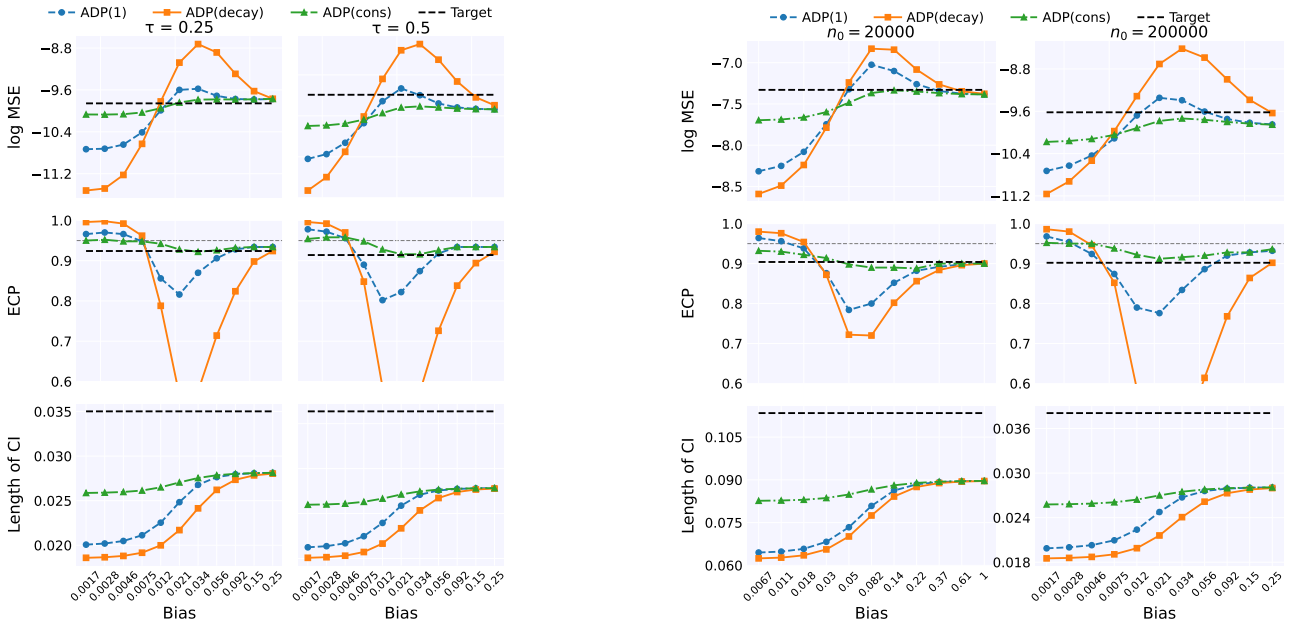


Figure A.3. MSEs on the log scale (↓), empirical coverage probabilities at the 95% nominal level, and averaged confidence interval lengths (↓) under continuously varying bias values. The left panel fixes the target sample size at $n_0 = 20{,}000$, and the right panel fixes the quantile level at $\tau = 0.75$. Data for both panels are generated from normal distributions, with the response rate fixed at $r_k = 0.5$ for all sites.

β	Method	0.007	0.011	0.018	0.030	0.135	0.223	0.368	0.607	1.000
0.65	ADP(1)	96.9	96.2	93.6	85.8	88.2	91.0	92.0	92.1	92.3
	ADP(cons)	94.0	94.2	93.8	93.2	91.2	91.7	92.0	92.2	92.4
	ADP(decay)	98.4	97.6	94.2	83.4	84.2	88.6	90.4	91.0	91.2
0.70	ADP(1)	96.1	95.6	94.0	87.5	87.4	90.6	91.8	91.8	91.7
	ADP(cons)	94.0	93.9	93.6	92.9	90.9	91.8	91.7	91.7	91.7
	ADP(decay)	98.0	97.8	94.4	86.6	81.8	87.6	89.8	91.0	91.6
0.75	ADP(1)	95.6	95.6	94.5	89.9	85.3	89.1	90.5	91.3	91.5
	ADP(cons)	93.8	93.6	93.6	93.0	90.0	90.2	90.6	91.3	91.4
	ADP(decay)	98.0	97.8	96.2	90.6	79.2	84.8	88.4	89.0	89.8

Table A.2. Empirical 95% coverage probabilities under bias values ranging from $\exp(-5)$ to 1, across varying learning rate parameters β. The target sample size is fixed at $n_0 = 20{,}000$, with the target quantile set to $\tau = 0.5$. Data are generated from normal distributions with a fixed response rate of $r_k = 0.5$ across all sites.

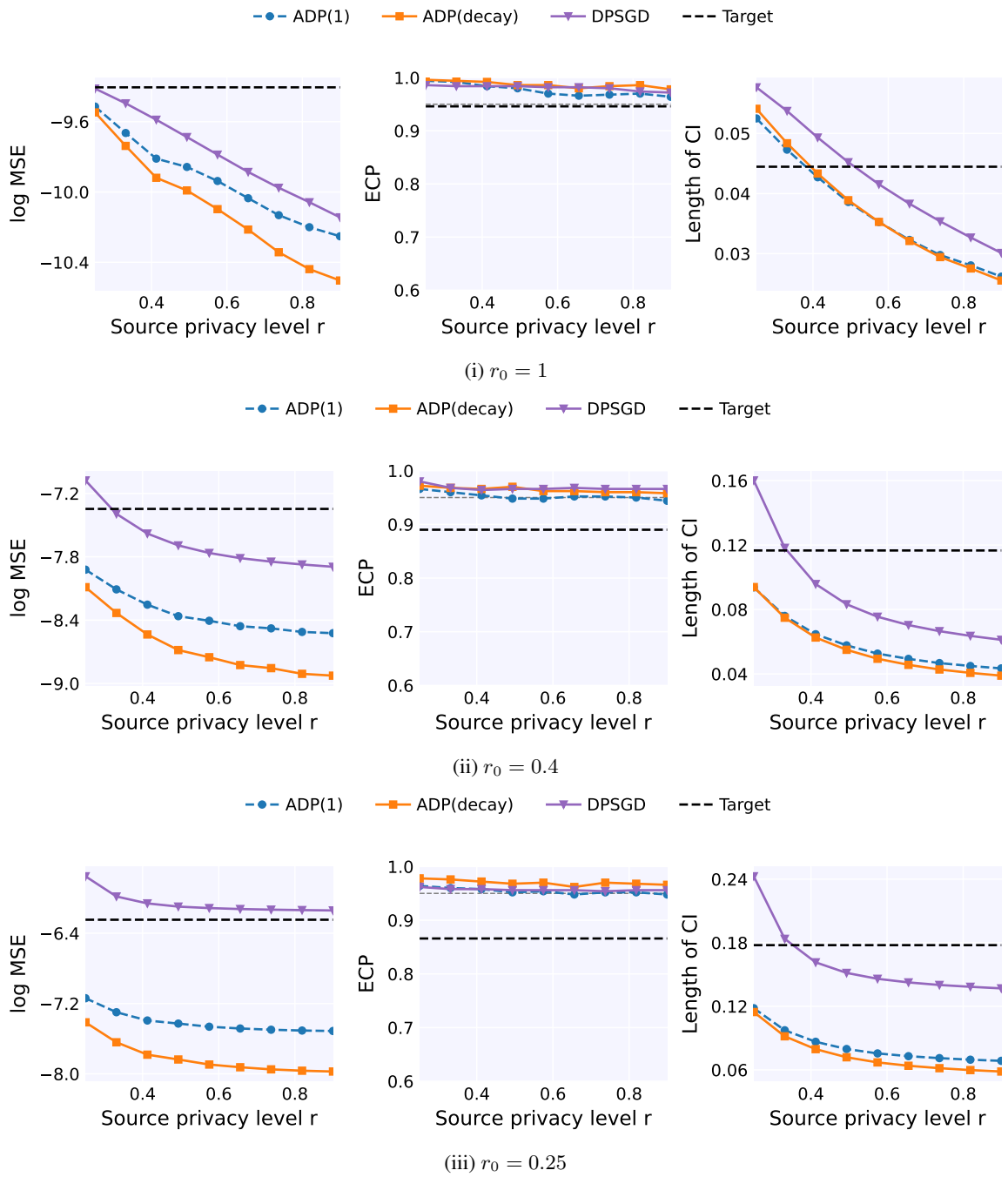


Figure A.4. MSEs on the log scale (↓), empirical coverage probabilities at the 95% nominal level, and averaged confidence interval lengths (↓) under varying truthful response rates $r_k = r$ for the source site. The target sample size is fixed at $n_0 = 20{,}000$, with the target quantile set to $\tau = 0.5$. Data are generated from normal distributions with a fixed $\mu_k = 0$ across all sites.

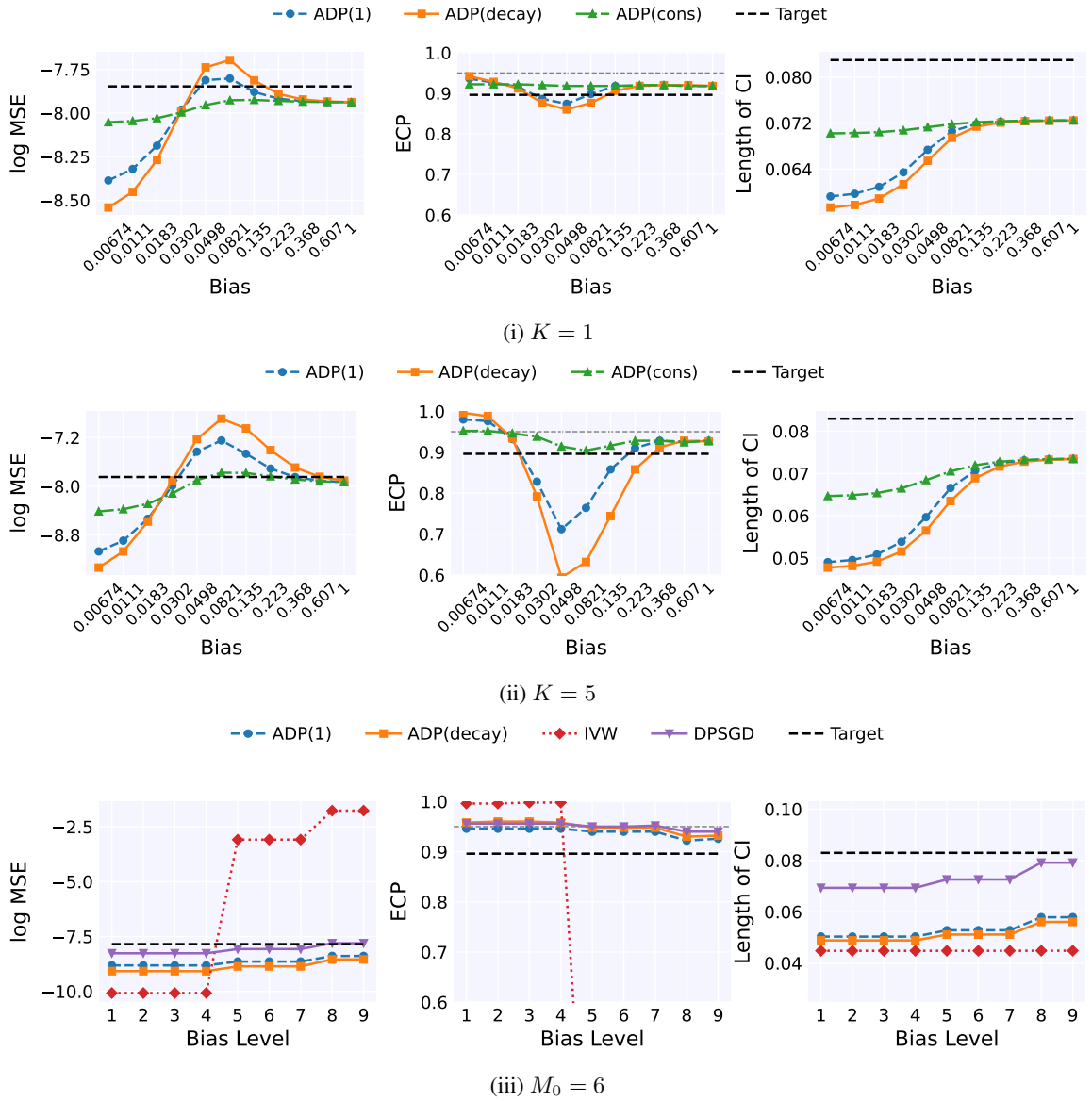
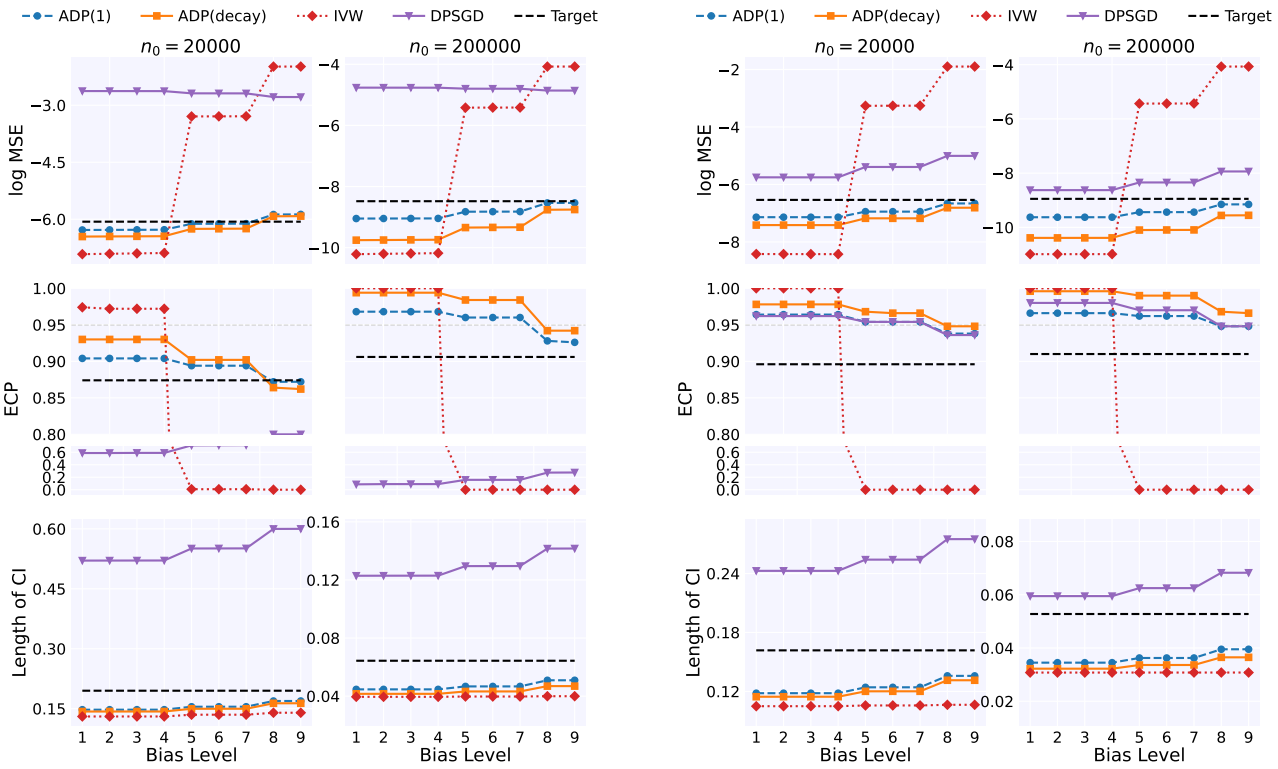
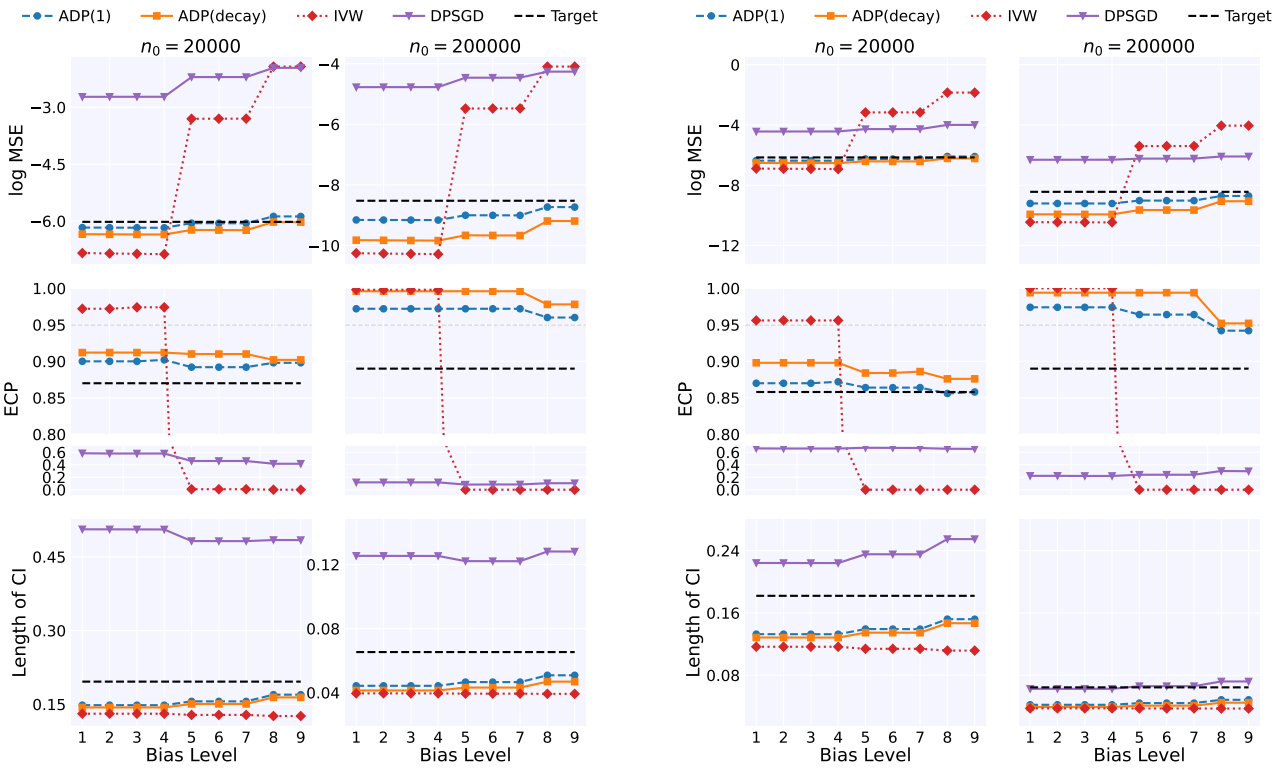


Figure A.5. MSEs on the log scale (↓), empirical coverage probabilities at the 95% nominal level, and averaged confidence-interval lengths (↓) under different heterogeneity scenarios. The target sample size is fixed at $n_0 = 20{,}000$, with the target quantile set to $\tau = 0.5$. Data are generated from Normal distributions with a fixed response rate of $r_k = 0.5$ across all sites.



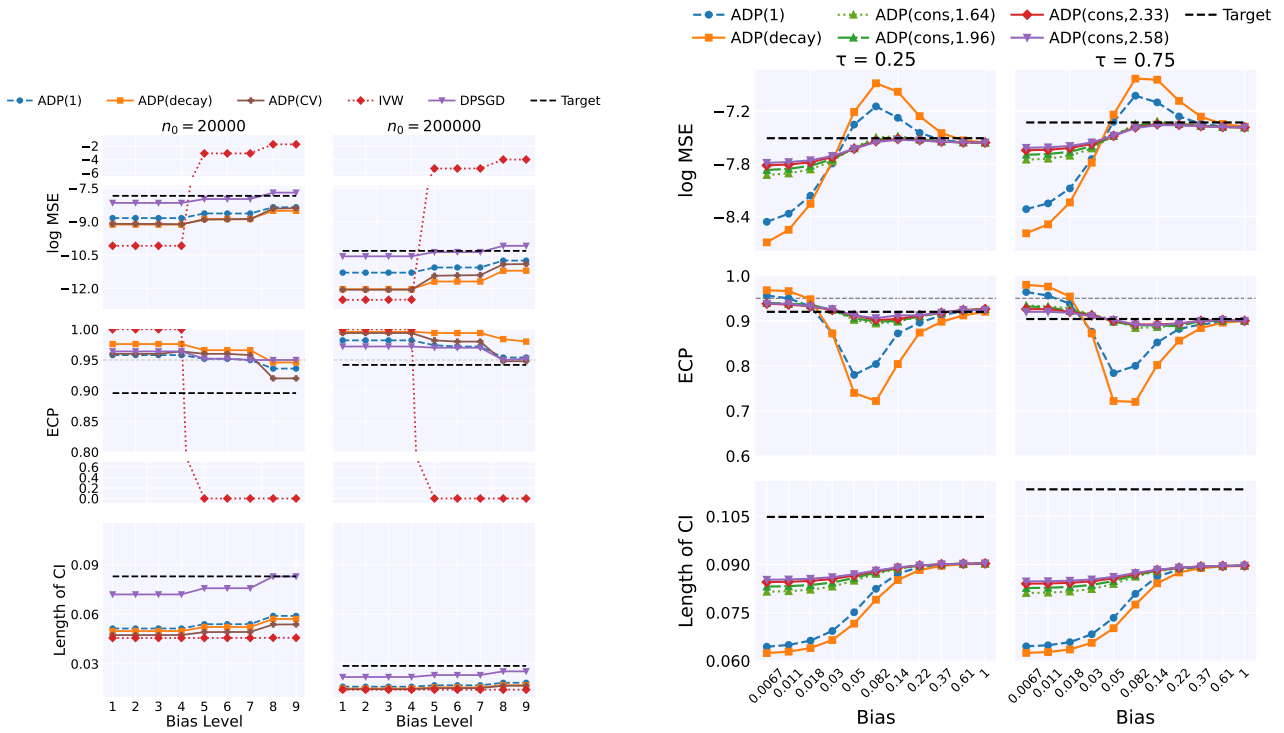
(i) $\tau = 0.25$ (ii) $\tau = 0.5$

Figure A.6. MSEs on the log scale (↓), empirical coverage probabilities at the 95% nominal level and the averaged confidence interval lengths (↓) under scenarios of either vanishing or distinguishable bias. Bias levels range from complete homogeneity (level 1) to strong heterogeneity (level 9). Data for both panels are generated from normal distributions, with the response rate fixed at $r_k = 0.25$ for all sites.



(i) $r_k = 0.25$ (ii) $r_k = 0.5$

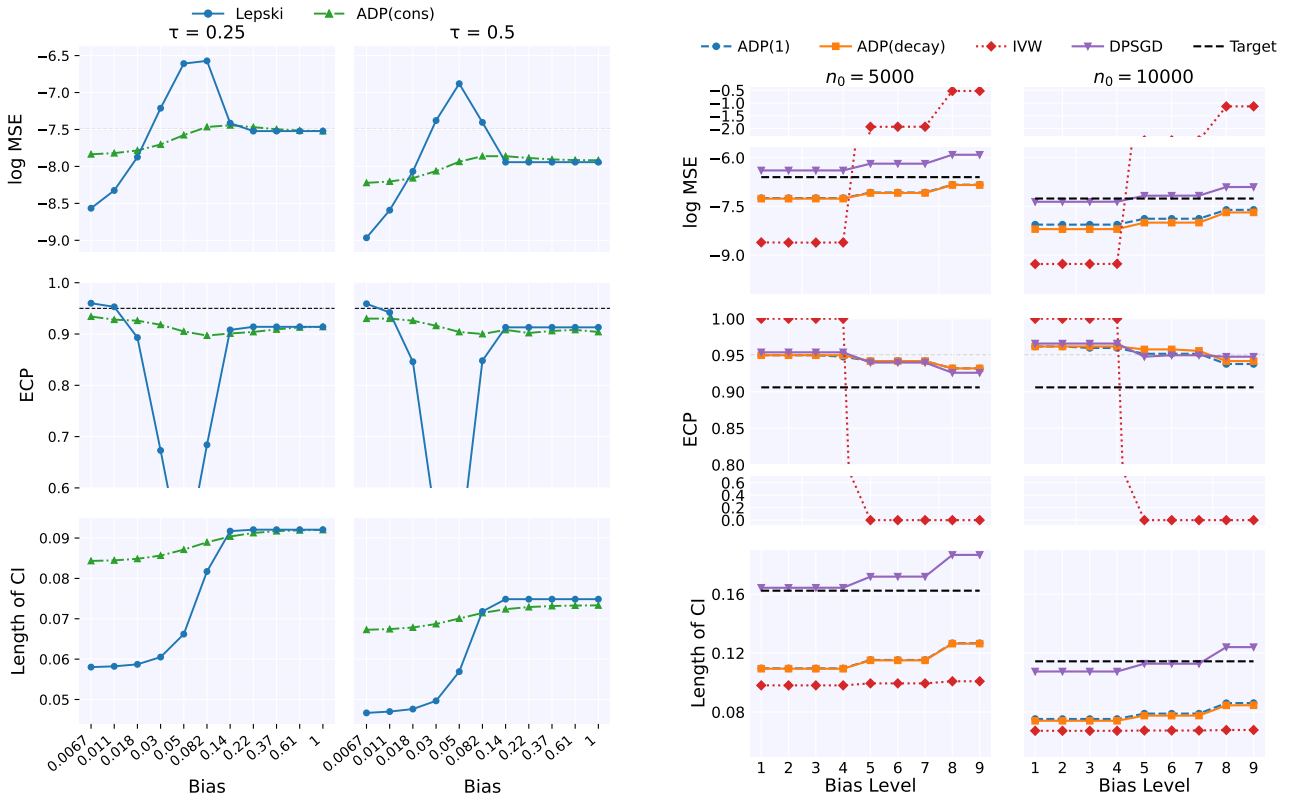
Figure A.7. MSEs on the log scale (↓), empirical coverage probabilities at the 95% nominal level and the averaged confidence interval lengths (↓) under scenarios of either vanishing or distinguishable bias. Bias levels range from complete homogeneity (level 1) to strong heterogeneity (level 9). Data for the left panels are generated from the normal distribution, and those for the right panels are generated from the Cauchy distribution.



(i) ADP(CV) for choosing λ

(ii) sensitivity to $\mathcal{C}$

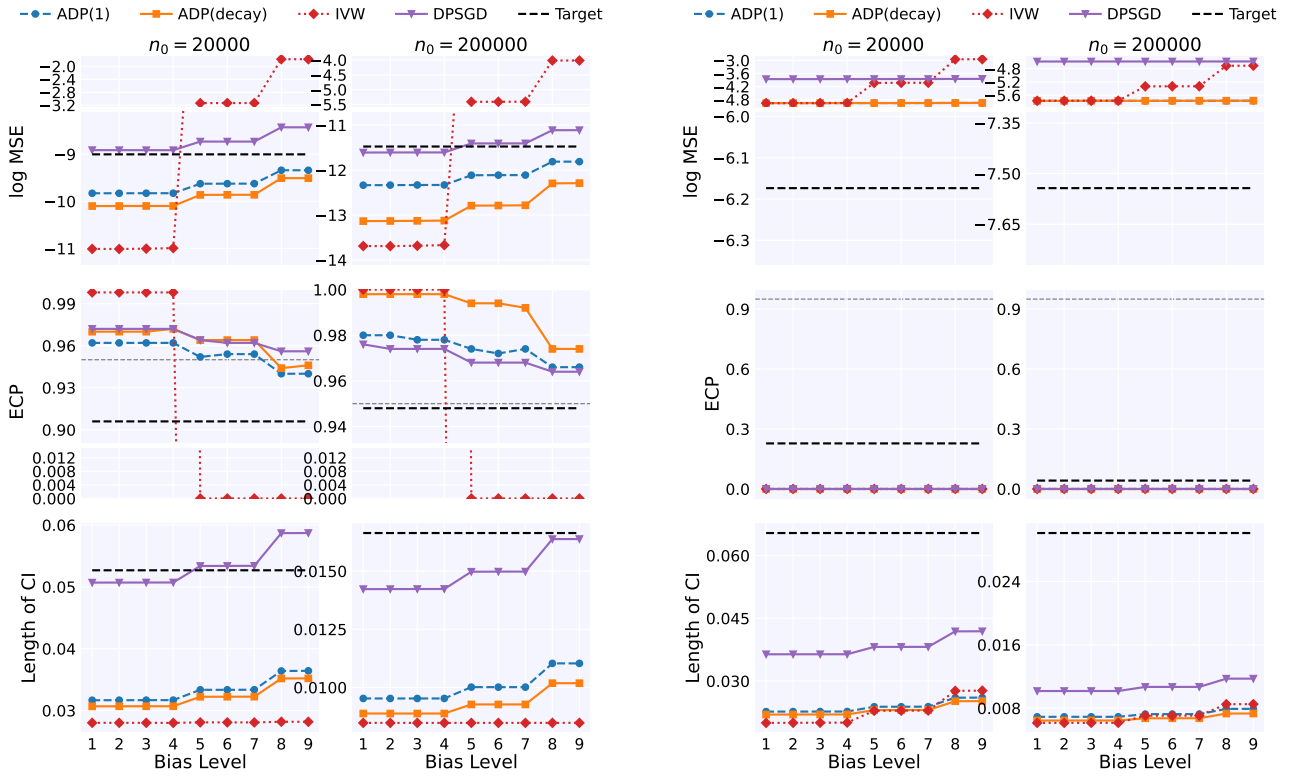
Figure A.8. Additional tuning-parameter analyses. Panel (i) compares ADP(CV) with the default choices for λ under vanishing or distinguishable bias. Panel (ii) examines sensitivity to $\mathcal{C} \in \{1.64, 1.96, 2.33, 2.58\}$ under continuously varying bias values. The response rate is fixed at $r_k = 0.5$ for all sites, and the quantile level is $\tau = 0.5$.



(i) Lepski-type baseline

(ii) smaller n_0

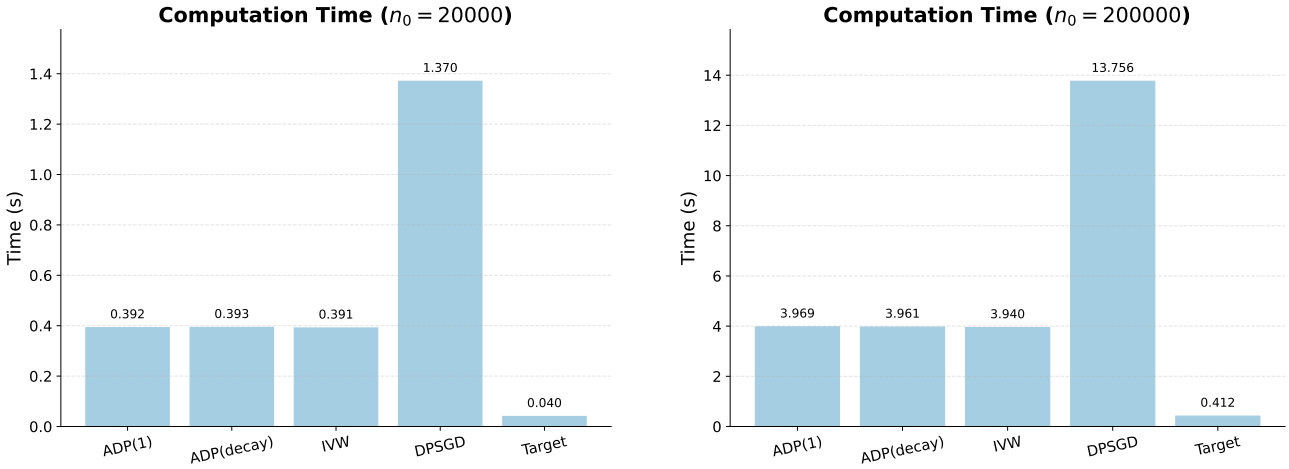
Figure A.9. Additional finite-sample and baseline comparisons. Panel (i) compares ADP(cons) with a simple Lepski-type hard-selection baseline under continuously varying bias values. Panel (ii) considers smaller target sample sizes $n_0 \in \{5{,}000, 10{,}000\}$ under the same setup as Section 5.1. The response rate is fixed at $r_k = 0.5$ for all sites, and $\tau = 0.5$.



(i) Laplace design

(ii) censored-normal design

Figure A.10. Robustness under assumption-violation designs. Panel (i) replaces the error distribution by $\mathrm{Laplace}(0, 1)$, violating the smoothness condition in Assumption 4.1; its qualitative conclusions are broadly consistent with the main vanishing/distinguishable-bias experiment. Panel (ii) censors latent normal observations at zero, introducing a point mass at the target quantile and violating continuity/absolute continuity; in this case, empirical coverage deteriorates substantially. All other settings follow Section 5.1, with $r_k = 0.5$ and $\tau = 0.5$.



(i) $n_0 = 20{,}000$

(ii) $n_0 = 200{,}000$

Figure A.11. Serial wall-clock time in seconds (↓) per replication for different methods. The response rate is $r_k = 0.5$ for all sites, the quantile level is $\tau = 0.5$, and we use the fifth bias level among the nine bias settings in Section 5.1.

B. Technical Lemmas

Notations. We first introduce some notation used throughout the paper. For two positive sequences a_n and b_n, write $a_n \ll b_n$ or $a_n = \mathrm{o}(b_n)$ if $a_n/b_n \to 0$ as $n \to \infty$. Similarly, write $a_n \lesssim b_n$ or $a_n = O(b_n)$ if there exists a constant $C < \infty$ such that $a_n/b_n \leq C$ for all sufficiently large n. Moreover, write $a_n \asymp b_n$ if both $a_n \lesssim b_n$ and $b_n \lesssim a_n$ hold simultaneously. For two random sequences $\{X_n\}$ and $\{Y_n\}$, write $X_n = \mathrm{o}_p(Y_n)$ if $P(|X_n/Y_n| > \varepsilon) \to 0$ for any $\varepsilon > 0$. Write $X_n = O_p(Y_n)$ if for every $\varepsilon > 0$, there exists a constant $M > 0$ such that $\limsup_{n\to\infty} P\big(|X_n|/|Y_n| > M\big) < \varepsilon$. Write $X_n \xrightarrow{d} X$ if the random sequence X_n converges in distribution to a random variable X. Finally, $\lfloor x \rfloor$ denotes the largest integer less than or equal to x.

Lemma B.1. *Denote $Z_k^{(m)} = \sqrt{(n_k/M_k)}(\widehat{\theta}_k^{(m)} - \theta_k)$. We have*

$$\mathbb{E}(Z_k^{(m)})^2 - \sigma_k^2 = \mathcal{O}\Big(1/(n_k/M_k)^{(\beta-1/2)\wedge(1-\beta)}\Big).$$

$$\mathbb{E}(Z_k^{(m)}) = \mathcal{O}\Big(1/(n_k/M_k)^{(\beta/2-1/4)\wedge(1/2-\beta/2)}\Big).$$

Proof. The proof is shown in Corollary 6 in (Gadat & Panloup, 2023). □

Lemma B.2. *For each $0 \leq k \leq K$, define the weighted estimator as $\widehat{\theta}(\mathbf{w}) = \sum_{k=0}^K w_k \widehat{\theta}_k$, let $\theta^*(\mathbf{w}) = \sum_{k=0}^K w_k \theta_k$ and $\sigma^2(\mathbf{w}) = \sum_{k=0}^K \frac{N}{n_k} w_k^2 \sigma_k^2$ represent its corresponding true parameter and variance, respectively. Then under assumption 4.7, we have:*

$$\sqrt{N}\left(\widehat{\theta}(\mathbf{w}) - \theta^*(\mathbf{w})\right) \xrightarrow{d} \mathcal{N}\left(0, \sigma^2(\mathbf{w})\right).$$

Proof. Lemma B.2 follows naturally as a corollary of Theorem 4.6. □

Lemma B.3. *For every $0 \leq k \leq K$, we have*

$$\mathbb{E}|\widehat{b}_k - b_k| \lesssim \frac{1}{\sqrt{N}} \tag{B.1}$$

$$\mathbb{E}|\widehat{b}_k^2 - b_k^2| \lesssim \frac{1}{N} + |b_k|\frac{1}{\sqrt{N}} \tag{B.2}$$

Proof. Denote $Z_k^{(m)} = \sqrt{(n_k/M_k)}(\widehat{\theta}_k^{(m)} - \theta_k)$, $\bar{Z}_k = M_k^{-1}\sum_{m=1}^{M_k} Z_k^{(m)}$ for every $0 \leq k \leq K$.

(i): Proof of equation (B.1). Recall that $b_k = \theta_k - \theta_0$ and $\widehat{b}_k = \widehat{\theta}_k - \widehat{\theta}_0$, with $\widehat{\theta}_k = \frac{1}{M_k}\sum_{m=1}^{M_k}\widehat{\theta}_k^{(m)}$. First, note that

$$\mathbb{E}|\widehat{b}_k - b_k| \lesssim \sqrt{\mathbb{E}|\widehat{b}_k - b_k|^2}.$$

It then suffices to analyze $\mathbb{E}|\widehat{b}_k - b_k|^2$. It could be proved that under Assumption 4.3

$$\begin{aligned}\mathbb{E}|\widehat{b}_k - b_k|^2 &\lesssim \mathbb{E}\big(\widehat{\theta}_0 - \theta_0\big)^2 + \mathbb{E}\big(\widehat{\theta}_k - \theta_k\big)^2 = \frac{M_0}{n_0}\big\{\mathbb{E}^2(\bar{Z}_0) + \mathrm{var}(\bar{Z}_0)\big\} \\ &\quad + \frac{M_k}{n_k}\big\{\mathbb{E}^2(\bar{Z}_k) + \mathrm{var}(\bar{Z}_k)\big\} \lesssim \frac{M_0}{n_0}\Big(1/(n_0/M_0)^{(\beta-1/2)\wedge(1-\beta)}\Big) \\ &\quad + \frac{M_k}{n_k}\Big(1/(n_k/M_k)^{(\beta-1/2)\wedge(1-\beta)}\Big) + \frac{1}{N} \lesssim \frac{1}{N}.\end{aligned}$$

This yields $\mathbb{E}|\widehat{b}_k - b_k| \lesssim 1/\sqrt{N}$.

(ii): Proof of equation (B.2). It could be verified that

$$\begin{aligned}\mathbb{E}|\widehat{b}_k^2 - b_k^2| &\leq \mathbb{E}|\widehat{b}_k - b_k|^2 + 2b_k\mathbb{E}|\widehat{b}_k - b_k| \lesssim \mathbb{E}|\widehat{b}_k - b_k|^2 + |b_k|\sqrt{\mathbb{E}|\widehat{b}_k - b_k|^2} \\ &\lesssim \frac{1}{N} + \frac{|b_k|}{\sqrt{N}}.\end{aligned}$$

This finishes the whole Lemma proof. □

Lemma B.4 (Consistency). *Under Assumption 4.7, consider the following three scenarios: (a) $\lambda = 0$, and the bias scale satisfies Assumption 4.10 (1); (b) λ is bounded away from 0, and the bias scale satisfies Assumption 4.10 (2); (c) λ satisfies Assumption 4.9, and the bias scale satisfies Assumption 4.10. Then, for each $0 \le k \le K$, with probability $1 - o(1)$, we have*

$$|\widehat{w}_k - w_k^*| \lesssim \left\{ \sum_{k=0}^{K} \prod_{j \neq k} \left(\frac{\sigma_j^2}{N} + \lambda b_j^2 \right) \right\}^{-1} \left\{ \sum_{k=0}^{K} \sum_{i \neq k} \left(\frac{\lambda |b_i|}{\sqrt{N}} + \frac{o(1)}{N} \right) \prod_{j \neq i}^{j \neq k} \left(\frac{\sigma_j^2}{N} + \lambda b_j^2 \right) \right\} = o(1).$$

C. Proof of the Main Theoretical Results

Without loss of generality, we simply assume that $\lfloor n_k / M_k \rfloor = n_k / M_k =: T_k$.

C.1. Proof of Theorem 4.6

First, we have $\mathbb{E}|\widehat{b}_k^2 - b_k^2| \lesssim 1/N + |b_k|/\sqrt{N}$ in Lemma B.3. Next, for any fixed $k = 0, \ldots, K$, one rewrites

$$\widehat{\theta}_k - \theta_k = \frac{1}{M_k} \sum_{m=1}^{M_k} \frac{1}{T_k} \sum_{t=1}^{T_k} (\widehat{\theta}_{k,t}^{(m)} - \theta_k) =: \frac{1}{M_k} \sum_{m=1}^{M_k} \mathcal{T}_k^{(m)},$$

where

$$\mathcal{T}_k^{(m)} = \widehat{\theta}_k^{(m)} - \theta_k = \frac{1}{T_k} \sum_{t=1}^{T_k} (\widehat{\theta}_{k,t}^{(m)} - \theta_k).$$

Let $U_{k,t}^{(m)}$ and $V_{k,t}^{(m)}$ be i.i.d. Bernoulli variables, mutually independent and also independent of $X_{k,t}^{(m)}$, with

$$\mathbb{P}(U_{k,t}^{(m)} = 1) = r_k, \quad \mathbb{P}(U_{k,t}^{(m)} = 0) = 1 - r_k, \qquad \mathbb{P}(V_{k,t}^{(m)} = 1) = \mathbb{P}(V_{k,t}^{(m)} = 0) = \tfrac{1}{2}.$$

Then it could be found that equation (6) could be rewritten as

$$\widehat{\theta}_{k,t+1}^{(m)} = \widehat{\theta}_{k,t}^{(m)} - \eta_{k,t} G_k(\widehat{\theta}_{k,t}^{(m)}, \zeta_{k,t+1}^{(m)}),$$

where $\zeta_{k,t}^{(m)} = \left(X_{k,t}^{(m)}, U_{k,t}^{(m)}, V_{k,t}^{(m)} \right)^\top$, and

$$\begin{aligned} G_k(\theta, \zeta_{k,t}^{(m)}) = & \frac{1 + r_k - 2r_k\tau}{2} \left\{ \mathbf{1}(X_{k,t}^{(m)} \le \theta) U_{k,t}^{(m)} + (1 - U_{k,t}^{(m)})(1 - V_{k,t}^{(m)}) \right\} \\ & - \frac{1 - r_k + 2r_k\tau}{2} \left\{ \mathbf{1}(X_{k,t}^{(m)} > \theta) U_{k,t}^{(m)} + (1 - U_{k,t}^{(m)}) V_{k,t}^{(m)} \right\}. \end{aligned}$$

Define

$$\varepsilon_{k,t}^{(m)} = g_k(\widehat{\theta}_{k,t-1}^{(m)}) - G_k(\widehat{\theta}_{k,t-1}^{(m)}, \zeta_{k,t}^{(m)}), \;\; \widetilde{\varepsilon}_{k,t}^{(m)} = g_k(\theta_k) - G_k(\theta_k, \zeta_{k,t}^{(m)}).$$

Elementary calculation shows that

$$\begin{aligned} \mathbb{E}(\varepsilon_{k,t}^{(m)2} | \mathcal{F}_{k,t-1}^{(m)}) &= \frac{1 + r_k - 2r_k F_k(\widehat{\theta}_{k,t-1}^{(m)})}{2} - \left(\frac{1 + r_k - 2r_k F_k(\widehat{\theta}_{k,t-1}^{(m)})}{2} \right)^2 \\ &= \frac{1 - r_k^2 \left\{ 2F_k(\widehat{\theta}_{k,t-1}^{(m)}) - 1 \right\}^2}{4} \\ &\xrightarrow{\mathbb{P}} \frac{1 - r_k^2 (2\tau - 1)^2}{4}, \end{aligned}$$

where the convergence in probability holds by the consistency of the quantile estimation and the continuous mapping theorem. Denote $\Delta_{k,t}^{(m)} = \widehat{\theta}_{k,t}^{(m)} - \theta_k$, $H_k = r_k f_k(\theta_k)$, $B_{k,t} = 1 - \eta_{k,t} H_k$, $A_{k,j}^t = \sum_{s=j}^t \left(\prod_{i=j+1}^s B_{k,i}\right) \eta_{k,i}$ for any $j \le t$. We decompose that

$$\begin{aligned}
\mathcal{T}_k^{(m)} &= \frac{1}{T_k}\sum_{j=1}^{T_k}(\widehat{\theta}_{k,j}^{(m)} - \theta_k) = \frac{1}{T_k}\sum_{j=1}^{T_k}\Delta_{k,j}^{(m)} \\
&= \frac{1}{T_k}A_{k,0}^{T_k-1}B_{k,0}\Delta_{k,0} + \frac{1}{T_k}\sum_{j=0}^{T_k-1}A_{k,j}^{T_k-1}r_{k,j}^{(m)} + \frac{1}{T_k}\sum_{j=0}^{T_k-1}\left(A_{k,j}^{T_k-1} - H_k^{-1}\right)\varepsilon_{k,j+1}^{(m)} \\
&+ \frac{1}{T_k}\sum_{j=0}^{T_k-1}H_k^{-1}\left(\varepsilon_{k,j+1}^{(m)} - \widetilde{\varepsilon}_{k,j+1}^{(m)}\right) + \frac{1}{T_k}\sum_{j=0}^{T_k-1}H_k^{-1}\widetilde{\varepsilon}_{k,j+1}^{(m)} \\
&=: \mathcal{T}_{k,1}^{(m)} + \mathcal{T}_{k,2}^{(m)} + \mathcal{T}_{k,3}^{(m)} + \mathcal{T}_{k,4}^{(m)} + \mathcal{T}_{k,5}^{(m)},
\end{aligned}$$

in which

$$r_{k,j}^{(m)} = H_k(\widehat{\theta}_{k,j}^{(m)} - \theta_k) - g_k(\widehat{\theta}_{k,j}^{(m)}).$$

For $\mathcal{T}_{k,1}^{(m)}$: According to Lemma C.4 of (Xie et al., 2024), one has $\left|A_{k,0}^{t-1}\right| \le C_0$ uniformly for all $t \ge 1$. Further observe that $\Delta_{k,0}^{(m)} \equiv \widehat{\theta}_{k,0} - \theta_k$ for all $1 \le m \le M_k$, thus one obtains

$$\left|\frac{1}{M_k}\sum_{m=1}^{M_k}\mathcal{T}_{k,1}^{(m)}\right| = \mathcal{O}\left(M_k/n_k\right).$$

For $\mathcal{T}_{k,2}^{(m)}$: Theorem 5 of (Gadat & Panloup, 2023) shows that

$$\max_{1\le m\le M_k}\mathbb{E}\left|\widehat{\theta}_{k,t}^{(m)} - \theta_k\right|^2 \lesssim \eta_{k,t}.$$

According to the uniform boundedness of $\left|A_{k,j}^{t-1}\right|$ and the fact that $|r_{k,t}^{(m)}| \lesssim |\widehat{\theta}_{k,t}^{(m)} - \theta_k|^2$,

$$\begin{aligned}
\mathbb{E}\left|\frac{1}{M_k}\sum_{m=1}^{M_k}\mathcal{T}_{k,2}^{(m)}\right| &\lesssim \mathbb{E}\frac{1}{M_k}\sum_{m=1}^{M_k}\frac{1}{T_k}\sum_{t=0}^{T_k-1}\left|r_{k,t}^{(m)}\right| \\
&\le \frac{1}{M_k}\sum_{m=1}^{M_k}\frac{1}{T_k}\sum_{t=0}^{T_k-1}\mathbb{E}|\widehat{\theta}_{k,t}^{(m)} - \theta_k|^2 \\
&= \mathcal{O}\left((M_k/n_k)^\beta\right).
\end{aligned}$$

Hence,

$$\left|\frac{1}{M_k}\sum_{m=1}^{M_k}\mathcal{T}_{k,2}^{(m)}\right| = \mathcal{O}_p\left((M_k/n_k)^\beta\right).$$

For $\mathcal{T}_{k,3}^{(m)}$: For any fixed $p > 0$, note that $\max_{1\le m\le M_k}\mathbb{E}|\varepsilon_{k,j}^{(m)}|^{2p} = \mathbb{E}|\varepsilon_{1,j}^{(m)}|^{2p}$ is bounded. Following the arguments in (Xie et al., 2024), one has $\left\|\mathcal{T}_{3,k}^{(m)}\right\|_{2p} = \mathcal{O}\left((n_k/M_k)^{-1+\beta/2}\right)$. Then, using the Lemma A in Chapter 9.2.6 of (Serfling, 2009) and the independence over m, one has that

$$\left\|\frac{1}{M_k}\sum_{m=1}^{M_k}\mathcal{T}_{k,3}^{(m)}\right\|_{2p} = \mathcal{O}\left(M_k^{-1/2}(n_k/M_k)^{-1+\beta/2}\right),$$

which implies that

$$\left|\frac{1}{M_k}\sum_{m=1}^{M_k}\mathcal{T}_{k,3}^{(m)}\right| = \mathcal{O}_p\left(M_k^{-1/2}(n_k/M_k)^{-1+\beta/2}\right).$$

For $\mathcal{T}_{k,4}^{(m)}$: Observe that $\varepsilon_{k,j}^{(m)} - \widetilde{\varepsilon}_{k,j}^{(m)} = g_k(\widehat{\theta}_{k,j-1}^{(m)}) - G_k(\widehat{\theta}_{k,j-1}^{(m)}, \zeta_{k,j}^{(m)}) + G_k(\theta_k, \zeta_{k,j}^{(m)})$, and

$$\begin{aligned}\mathbb{E}\left(\varepsilon_{k,j}^{(m)} - \widetilde{\varepsilon}_{k,j}^{(m)}\right)^2 &\lesssim \mathbb{E}g_k^2(\widehat{\theta}_{k,j-1}^{(m)}) + \mathbb{E}\left\{G_k(\widehat{\theta}_{k,j-1}^{(m)}, \zeta_{k,j}) - G_k(\theta_k, \zeta_{k,j})\right\}^2 \\ &\lesssim \mathbb{E}\left|\widehat{\theta}_{k,j-1}^{(m)} - \theta_k\right|^2 + \mathbb{E}\left|\widehat{\theta}_{k,j-1}^{(m)} - \theta_k\right| \\ &\lesssim \mathbb{E}\left|\widehat{\theta}_{k,j-1}^{(m)} - \theta_k\right|^2 + \left\{\mathbb{E}\left|\widehat{\theta}_{k,j-1}^{(m)} - \theta_k\right|^2\right\}^{1/2} \lesssim \eta_{k,t}^{1/2},\end{aligned}$$

where the last inequality holds by Theorem 5 of (Gadat & Panloup, 2023), and the constant does not depend on m.

Observe that $\sum_{t=1}^{T_k} H_k^{-1}\left(\varepsilon_{k,t}^{(m)} - \widetilde{\varepsilon}_{k,t}^{(m)}\right)$ is a martingale for each m (independent over $1 \le m \le M_k$), Burkholder's inequality entails that

$$\begin{aligned}\left\|\sum_{t=1}^{T_k} H_k^{-1}\left(\varepsilon_{k,t}^{(m)} - \widetilde{\varepsilon}_{k,t}^{(m)}\right)\right\|_2 &\lesssim \left\{\sum_{t=1}^{T_k}\left\|H_k^{-1}\left(\varepsilon_{k,t}^{(m)} - \widetilde{\varepsilon}_{k,t}^{(m)}\right)\right\|_2^2\right\}^{1/2} \\ &\lesssim \left(\sum_{t=1}^{T_k}\eta_{k,t}^{1/2}\right)^{1/2} \lesssim (n_k/M_k)^{\{1-\beta/2\}/2}.\end{aligned}$$

Hence,

$$\left\|\mathcal{T}_{k,4}^{(m)}\right\|_2 = \left\|\frac{1}{T_k}\sum_{j=1}^{T_k} H^{-1}\left(\varepsilon_{k,j}^{(m)} - \widetilde{\varepsilon}_{k,j}^{(m)}\right)\right\|_2 = \mathcal{O}\left((n_k/M_k)^{-1/2-\beta/4}\right),$$

which further implies that

$$\frac{1}{M_k}\sum_{m=1}^{M_k}\mathcal{T}_{k,4}^{(m)} = \mathcal{O}_p\left(M_k^{-1/2}(n_k/M_k)^{-1/2-\beta/4}\right).$$

For $\mathcal{T}_{k,5}^{(m)}$: Elementary calculation shows that

$$\mathbb{E}\widetilde{\varepsilon}_{k,j}^{(m)2} = \frac{1 - r_k^2(2\tau-1)^2}{4} =: S_k.$$

Applying Theorem 2.6.7 of (Csörgo & Révész, 1981) with $H(x) = x^{2p}$ and $x_n = n_k^{\beta_0}$, there exist i.i.d. standard normal $\widetilde{Z}_{k,i}$'s and some $a_k, C_k > 0$ (depending on the distribution of $H_k^{-1}\widetilde{\varepsilon}_{k,j}^{(m)}$) such that

$$\mathbb{P}\left(\left|\sum_{m=1}^{M_k}\sum_{t=1}^{T_k}\frac{H_k^{-1}\widetilde{\varepsilon}_{k,t}^{(m)}}{\sqrt{H_k^{-1}S_kH_k^{-1}}} - \sum_{i=1}^{n_k}\widetilde{Z}_{k,i}\right| > n_k^{\beta_0}\right) \le C_k a_k^{-2p} n_k^{1-2p\beta_0}.$$

Thus,

$$\mathbb{P}\left(\left|\frac{1}{n_k}\sum_{m=1}^{M_k}\sum_{t=1}^{T_k}\frac{H_k^{-1}\widetilde{\varepsilon}_{k,t}^{(m)}}{\sqrt{H_k^{-1}S_kH_k^{-1}}} - \frac{1}{n_k}\sum_{i=1}^{n_k}\widetilde{Z}_{k,i}\right| > n_k^{-1+\beta_0}\right) \lesssim n_k^{1-2p\beta_0}.$$

For $p > 2$, one selects $\beta_0 \in (1/p, 1/2)$, the Borel-Cantelli lemma leads to

$$\left| \frac{1}{M_k} \sum_{m=1}^{M_k} \mathcal{T}_{k,5}^{(m)} - \frac{1}{n_k} \sum_{i=1}^{n_k} Z_{k,i} \right| = \mathcal{O}_{a.s.} \left(n_k^{-1+\beta_0} \right),$$

where $Z_{k,i}$'s are i.i.d. normal r.v.'s with mean zero and covariance $H_k^{-1} S_k H_k^{-1}$.

Therefore, we obtain that

$$\left| \widehat{\theta}_k - \theta_k - \frac{1}{n_k} \sum_{i=1}^{n_k} Z_{k,i} \right| = o_p \left(n_k^{-1/2} \right),$$

which completes the proof of the weak convergence result.

As for the consistency of $\widehat{\sigma}_k^2$, we rewrite

$$\begin{aligned} \widehat{\sigma}_k^2 &= \frac{n_k}{M_k - 1} \frac{1}{M_k} \sum_{m=1}^{M_k} \left(\widehat{\theta}_k^{(m)} - \theta_k \right)^2 - \frac{n_k}{M_k - 1} \left\{ \frac{1}{M_k} \sum_{m=1}^{M_k} \left(\widehat{\theta}_k^{(m)} - \theta_k \right) \right\}^2 \\ &= \frac{n_k}{M_k - 1} \frac{1}{M_k} \sum_{m=1}^{M_k} \mathcal{T}_k^{(m)2} - \frac{n_k}{M_k - 1} \left\{ \frac{1}{M_k} \sum_{m=1}^{M_k} \mathcal{T}_k^{(m)} \right\}^2. \end{aligned}$$

Recall the definitions of $\mathcal{T}_{k,j}^{(m)}$ for $1 \le j \le 5$.

For $\mathcal{T}_{k,1}^{(m)}$: According to Lemma C.4 of (Xie et al., 2024), one has $\left| A_{k,0}^{t-1} \right| \le C_0$ uniformly for all $t \ge 1$. Further observe that $\Delta_{k,0}^{(m)} \equiv \widehat{\theta}_{k,0} - \theta_k$ for all $1 \le m \le M_k$, thus one obtains $\left\| \mathcal{T}_{k,1}^{(m)} \right\|_2 = \mathcal{O}\left(T_k^{-1} \right)$, where the constant does not depend on m.

For $\mathcal{T}_{k,2}^{(m)}$: Consider that

$$\left| \frac{1}{T_k} \sum_{j=0}^{T_k - 1} A_{k,j}^{T_k - 1} r_{k,j}^{(m)} \right| \lesssim \frac{1}{T_k} \sum_{j=0}^{T_k - 1} \left| r_{k,j}^{(m)} \right| \lesssim \frac{1}{T_k} \sum_{j=0}^{T_k - 1} \left| \widehat{\theta}_{k,j}^{(m)} - \theta_k \right|^2.$$

Then,

$$\left\| \frac{1}{T_k} \sum_{j=0}^{T_k - 1} A_{k,j}^{T_k - 1} r_{k,j}^{(m)} \right\|_2 \lesssim \frac{1}{T_k} \sum_{j=0}^{T_k - 1} \left\| \widehat{\theta}_{k,j}^{(m)} - \theta_k \right\|_4^2.$$

Applying Theorem 5 of (Gadat & Panloup, 2023), we have

$$\mathbb{E} \left| \widehat{\theta}_{k,j}^{(m)} - \theta_k \right|^4 \lesssim \eta_{k,j}^2.$$

Since $\eta_{k,j} \asymp j^{-\beta}$ with $\beta > 1/2$, it follows that $\left\| \mathcal{T}_{k,2}^{(m)} \right\|_2 = o\left(T_k^{-1/2} \right)$.

For $\mathcal{T}_{k,3}$: As shown in the previous arguments, one has $\left\| \mathcal{T}_{3,k}^{(m)} \right\|_{2p} = \mathcal{O}\left(T_k^{-1+\beta/2} \right) = o\left(T_k^{-1/2} \right)$, since $\beta < 1$.

For $\mathcal{T}_{k,4}$: Observe that $\sum_{j=1}^{T_k} H_k^{-1} \left(\varepsilon_{k,j}^{(m)} - \widetilde{\varepsilon}_{k,j}^{(m)} \right)$ is a martingale for each m (independent over $1 \le m \le M_k$), Burkholder's inequality entails that

$$\begin{aligned} \left\| \sum_{j=1}^{T_k} H_k^{-1} \left(\varepsilon_{k,j}^{(m)} - \widetilde{\varepsilon}_{k,j}^{(m)} \right) \right\|_2 &\lesssim \left\{ \sum_{j=1}^{T_k} \left\| H_k^{-1} \left(\varepsilon_{k,j}^{(m)} - \widetilde{\varepsilon}_{k,j}^{(m)} \right) \right\|_2^2 \right\}^{1/2} \\ &\lesssim \left(\sum_{j=1}^{T_k} \eta_{k,j}^{1/2} \right)^{1/2} \lesssim T_k^{\{1-\beta/2\}/2}. \end{aligned}$$

Hence, for any $1 \le m \le M_k$,

$$\left\|\mathcal{T}_{k,4}^{(m)}\right\|_2 = \left\|\frac{1}{T_k}\sum_{j=1}^{T_k} H_k^{-1}\left(\varepsilon_{k,j}^{(m)} - \widetilde{\varepsilon}_{k,j}^{(m)}\right)\right\|_2 = \mathcal{O}\left(T_k^{-1/2-\beta/4}\right).$$

For $\mathcal{T}_{k,5}$: Applying Theorem 2.6.7 of (Csörgo & Révész, 1981) with $H(x) = x^{2p}$ and $x_n = vT_k^{\beta_0}$, there exist i.i.d. standard normal $\widetilde{Z}_{k,j}^{(m)}$'s and some $a_k, C_k > 0$ (depending on the distribution of $H_k^{-1}\widetilde{\varepsilon}_{k,j}^{(m)}$) such that

$$\mathbb{P}\left(\left|\sum_{j=1}^{T_k}\frac{H_k^{-1}\widetilde{\varepsilon}_{k,j}^{(m)}}{\sqrt{H_k^{-1}S_kH_k^{-1}}} - \sum_{j=1}^{T_k}\widetilde{Z}_{k,j}^{(m)}\right| > vT_k^{\beta_0}\right) \le C_k a_k^{-2p} v^{-2p} T_k^{1-2p\beta_0}.$$

Thus,

$$\mathbb{P}\left(\left|\frac{1}{T_k}\sum_{j=1}^{T_k}\frac{H_k^{-1}\widetilde{\varepsilon}_{k,j}^{(m)}}{\sqrt{H_k^{-1}S_kH_k^{-1}}} - \frac{1}{T_k}\sum_{j=1}^{T_k}\widetilde{Z}_{k,j}^{(m)}\right| > vT_k^{-1+\beta_0}\right) \lesssim v^{-2p}T_k^{1-2p\beta_0}.$$

Since $\mathbb{E}X^p = \int_0^\infty pv^{p-1}\,\mathbb{P}(|X| > v)dv$, we also have

$$\left\|\mathcal{T}_{k,5}^{(m)} - \frac{1}{T_k}\sum_{j=1}^{T_k} Z_{k,j}^{(m)}\right\|_2 = \mathcal{O}\left(T_k^{-1+\beta_0}\right).$$

According to the above results, we show that

$$\left\|\mathcal{T}_k^{(m)} - \frac{1}{T_k}\sum_{j=1}^{T_k} Z_{k,j}\right\|_2 = o\left(T_k^{-1/2}\right),$$

which implies

$$\left\|\mathcal{T}_k^{(m)}\right\|_2 = \left\|\frac{1}{\sqrt{T_k}}\sum_{j=1}^{T_k} Z_{k,j}\right\|_2 + o(1) < \infty.$$

The SLLN (i.i.d.) and the continuous mapping theorem further yields that

$$\frac{1}{M_k}\sum_{m=1}^{M_k}\left(\sqrt{T_k}\mathcal{T}_k^{(m)}\right)^2 \xrightarrow{a.s.} \sigma_k^2, \quad \left(\frac{1}{M_k}\sum_{m=1}^{M_k}\sqrt{T_k}\mathcal{T}_k^{(m)}\right)^2 \xrightarrow{a.s.} 0,$$

which completes the proof of consistency of $\widehat{\sigma}_k^2$.

C.2. Proof of Lemma B.4

Under Assumption 4.7, it is easy to obtain that for every $0 \le k \le K$,

$$|\widehat{b}_k - b_k| = \mathcal{O}_p\Big(\frac{1}{\sqrt{N}}\Big); \quad |\widehat{b}_k^2 - b_k^2| = \mathcal{O}_p\Big(\frac{1}{N}\Big) + |b_k|\mathcal{O}_p\Big(\frac{1}{\sqrt{N}}\Big). \tag{C.1}$$

Next, define the notation $\lesssim_p$ as follows: $a \lesssim_p b$ indicates that $a \lesssim b$ holds with probability $1 - o(1)$. Recall the definitions of w_k^* and $\widehat{w}_k$. Define $\Delta_k = \frac{\sigma_k^2}{n_k} + \lambda b_k^2$, and $\widehat{\Delta}_k = \frac{\widehat{\sigma}_k^2}{n_k} + \lambda \widehat{b}_k^2$. Then, we have

$$w_k^* = \frac{\Delta_k^{-1}}{\sum_{j=0}^K \Delta_j^{-1}}, \quad \widehat{w}_k = \frac{\widehat{\Delta}_k^{-1}}{\sum_{j=0}^K \widehat{\Delta}_j^{-1}}.$$

We first consider the case $K = 2$ and scenario (c), i.e., λ satisfies Assumption 4.9, and the bias scale satisfies Assumption 4.10. In this case, w_k^* and $\widehat{w}_k$ can be rewritten as

$$w_k^* = \frac{\Delta_{k_1}\Delta_{k_2}}{\Delta_0\Delta_1 + \Delta_0\Delta_2 + \Delta_1\Delta_2} = \frac{\Delta_{k_1}\Delta_{k_2}}{\sum_{k=0}^2 \prod_{j\neq k}\Delta_j}, \widehat{w}_k = \frac{\widehat{\Delta}_{k_1}\widehat{\Delta}_{k_2}}{\sum_{k=0}^2 \prod_{j\neq k}\widehat{\Delta}_j} \text{ for } k_1 \neq k_2, k_1 \neq k, k_2 \neq k.$$

It can then be verified that

$$\begin{aligned}
|\widehat{w}_k - w_k^*| &= \left|\frac{\Delta_{k_1}\Delta_{k_2}}{\sum_{k=0}^2 \prod_{j\neq k}\Delta_j} - \frac{\widehat{\Delta}_{k_1}\widehat{\Delta}_{k_2}}{\sum_{k=0}^2 \prod_{j\neq k}\widehat{\Delta}_j}\right| \\
&= \frac{\left|(\widehat{\Delta}_{k_1}\widehat{\Delta}_{k_2} - \Delta_{k_1}\Delta_{k_2})(\sum_{k=0}^2 \prod_{j\neq k}\Delta_j) + \Delta_{k_1}\Delta_{k_2}\{\sum_{k=0}^2(\prod_{j\neq k}\Delta_j - \prod_{j\neq k}\widehat{\Delta}_j)\}\right|}{(\sum_{k=0}^2 \prod_{j\neq k}\widehat{\Delta}_j)(\sum_{k=0}^2 \prod_{j\neq k}\widehat{\Delta}_j)} \\
&\leq \frac{\left|\widehat{\Delta}_{k_1}\widehat{\Delta}_{k_2} - \Delta_{k_1}\Delta_{k_2}\right| + \left|\sum_{k=0}^2(\prod_{j\neq k}\Delta_j - \prod_{j\neq k}\widehat{\Delta}_j)\right|}{\sum_{k=0}^2 \prod_{j\neq k}\Delta_j} \frac{\sum_{k=0}^2 \prod_{j\neq k}\Delta_j}{\sum_{k=0}^2 \prod_{j\neq k}\widehat{\Delta}_j}.
\end{aligned}$$

The inequality holds because $\Delta_{k_1}\Delta_{k_2} \leq \sum_{k=0}^2 \prod_{j\neq k}\Delta_j$. Under assumption 4.7, we know that $\widehat{\Delta}_j$ is a consistent estimator of Δ_j. Therefore, we obtain

$$|\widehat{w}_k - w_k^*| \lesssim_p \frac{\left|\widehat{\Delta}_{k_1}\widehat{\Delta}_{k_2} - \Delta_{k_1}\Delta_{k_2}\right| + \left|\sum_{k=0}^2(\prod_{j\neq k}\Delta_j - \prod_{j\neq k}\widehat{\Delta}_j)\right|}{\sum_{k=0}^2 \prod_{j\neq k}\Delta_j} \lesssim \frac{\sum_{k=0}^2 \left|\prod_{j\neq k}\widehat{\Delta}_j - \prod_{j\neq k}\Delta_j\right|}{\sum_{k=0}^2 \prod_{j\neq k}\Delta_j}. \tag{C.2}$$

We now analyze the numerator in equation (C.2). Under assumption 4.7 and equation (C.1), we have

$$\left|\widehat{\Delta}_k - \Delta_k\right| \lesssim \frac{|\widehat{\sigma}_k^2 - \sigma_k^2|}{N} + \lambda|\widehat{b}_k^2 - b_k^2| \lesssim_p \frac{o(1)}{N} + \frac{\lambda|b_k|}{\sqrt{N}}. \tag{C.3}$$

It can thus be shown that

$$\begin{aligned}
&\left|\widehat{\Delta}_{k_1}\widehat{\Delta}_{k_2} - \Delta_{k_1}\Delta_{k_2}\right| \leq \left|\Delta_{k_1}(\Delta_{k_2} - \widehat{\Delta}_{k_2})\right| + \left|\Delta_{k_2}(\Delta_{k_1} - \widehat{\Delta}_{k_1})\right| + \left|(\widehat{\Delta}_{k_2} - \Delta_{k_2})(\Delta_{k_1} - \widehat{\Delta}_{k_1})\right| \\
&\lesssim_p \left(\frac{\sigma_{k_1}^2}{N} + \lambda b_{k_1}^2\right)\left\{\frac{\lambda|b_{k_2}|}{\sqrt{N}} + \frac{o(1)}{N}\right\} + \left(\frac{\sigma_{k_2}^2}{N} + \lambda b_{k_2}^2\right)\left\{\frac{\lambda|b_{k_1}|}{\sqrt{N}} + \frac{o(1)}{N}\right\} + \left\{\frac{\lambda|b_{k_1}|}{\sqrt{N}} + \frac{o(1)}{N}\right\}\times \\
&\qquad \left\{\frac{\lambda|b_{k_2}|}{\sqrt{N}} + \frac{o(1)}{N}\right\} = \mathcal{E}_1 + \mathcal{E}_2 + \mathcal{E}_3.
\end{aligned}$$

If $|b_{k_1}| \ll N^{-1/2}$, we have

$$\frac{\lambda|b_{k_1}|}{\sqrt{N}} + \frac{o(1)}{N} \ll \frac{\sigma_{k_1}^2}{N} + \lambda b_{k_1}^2,$$

which implies $\mathcal{E}_3 \ll \mathcal{E}_1$. Similarly, if $|b_{k_2}| \ll N^{-1/2}$, we have $\mathcal{E}_3 \ll \mathcal{E}_2$. If both $\lambda|b_{k_1}| \gg N^{-1/2}$ and $\lambda|b_{k_2}| \gg N^{-1/2}$, then

$$\frac{\mathcal{E}_3}{\mathcal{E}_1} = \mathcal{O}\Big(\frac{\lambda^2|b_{k_1}||b_{k_2}|}{N} \times \frac{\sqrt{N}}{\lambda b_{k_1}^2\lambda|b_{k_2}|}\Big) = \mathcal{O}\Big(\frac{1}{\sqrt{N}|b_{k_1}|}\Big) = o(1).$$

Thus, we simplify to obtain

$$\begin{aligned}
\left|\widehat{\Delta}_{k_1}\widehat{\Delta}_{k_2} - \Delta_{k_1}\Delta_{k_2}\right| &\lesssim_p \left|\Delta_{k_1}(\widehat{\Delta}_{k_2} - \Delta_{k_2})\right| + \left|\Delta_{k_2}(\widehat{\Delta}_{k_1} - \Delta_{k_1})\right| \\
&\lesssim_p \left(\frac{\sigma_{k_1}^2}{N} + \lambda b_{k_1}^2\right)\left\{\frac{\lambda|b_{k_2}|}{\sqrt{N}} + \frac{o(1)}{N}\right\} + \left(\frac{\sigma_{k_2}^2}{N} + \lambda b_{k_2}^2\right)\left\{\frac{\lambda|b_{k_1}|}{\sqrt{N}} + \frac{o(1)}{N}\right\}.
\end{aligned}$$

Next, we analyze the denominator in equation (C.2). By definition, it could be shown that

$$\sum_{k=0}^2 \prod_{j\neq k}\Delta_j \asymp \sum_{k=0}^2 \prod_{j\neq k}\left(\frac{\sigma_j^2}{N} + \lambda b_j^2\right).$$

Substituting this result back into equation (C.2), we obtain

$$|\widehat{w}_k - w_k^*| \lesssim_p \left\{ \sum_{k=0}^{2} \prod_{j\neq k} \Big(\frac{\sigma_j^2}{N} + \lambda b_j^2 \Big) \right\}^{-1} \left\{ \sum_{k=0}^{2} \sum_{i\neq k} \Big(\frac{\lambda |b_i|}{\sqrt{N}} + \frac{o(1)}{N} \Big) \prod_{j\neq i}^{j\neq k} \Big(\frac{\sigma_j^2}{N} + \lambda b_j^2 \Big) \right\}.$$

We now consider different cases. First, note that we have $b_0 = 0$. For $k = 1, 2$, we discuss the following scenarios: (1) When $b_k \ll N^{-1/2}$ for $k = 1, 2$, the denominator is of order $1/N^2$, while the numerator is of order $o(1)/N^2$. (2) When $\lambda b_k \gg N^{-1/2}$ for one particular k and $b_j \ll N^{-1/2}$ for $j \neq k$, the denominator has order $\frac{\lambda b_k^2}{N}$, while the numerator has order $(\lambda b_k^2)o(1)/N + (\lambda |b_k|)/(N^{3/2})$. (3) When $\lambda b_k \gg N^{-1/2}$ for both $k = 1, 2$, the denominator has order $(\lambda b_1^2)(\lambda b_2^2)$, while the numerator is of order $(\lambda b_1^2)(\lambda |b_2|/\sqrt{N}) + (\lambda b_2^2)(\lambda |b_1|/\sqrt{N})$. In each scenario above, we can verify that $|\widehat{w}_k - w_k^*| = o_p(1)$.

For the general case with arbitrary K, the proof follows analogously. The detailed argument is omitted here, but it similarly leads to the result:

$$|\widehat{w}_k - w_k^*| \lesssim_p \left\{ \sum_{k=0}^{K} \prod_{j\neq k} \Big(\frac{\sigma_j^2}{N} + \lambda b_j^2 \Big) \right\}^{-1} \left\{ \sum_{k=0}^{K} \sum_{i\neq k} \Big(\frac{\lambda |b_i|}{\sqrt{N}} + \frac{o(1)}{N} \Big) \prod_{j\neq i}^{j\neq k} \Big(\frac{\sigma_j^2}{N} + \lambda b_j^2 \Big) \right\}.$$

Under Assumption 4.10, we therefore conclude that $|\widehat{w}_k - w_k^*| = o_p(1)$. Finally, noting that scenarios (a) and (b) are in fact special cases of scenario (c), we complete the proof of the theorem.

C.3. Proof of Theorem 4.12

The proof of the consistency of $\widehat{w}_k$ is provided in Appendix C.2. Next, we establish the asymptotic properties of the proposed estimator $\widehat{\theta}_{\text{est}}$. The proof is divided into four parts. In the first part, we prove that:

$$\sqrt{N}(\widehat{\theta}(\mathbf{w}^*) - \theta_0) \to_d \mathcal{N}\big(0, \sigma^2(\mathbf{w}^*)\big).$$

In the second part, we establish that:

$$\sqrt{N}(\widehat{\theta}(\mathbf{w}^*) - \widehat{\theta}_{\text{est}}) = o_p(1).$$

Combining the results from the first two parts immediately yields:

$$\sqrt{N}(\widehat{\theta}_{\text{est}} - \theta_0) \to_d \mathcal{N}\big(0, \sigma^2(\mathbf{w}^*)\big).$$

In the third part, we prove:

$$\widehat{\sigma}^2_{\text{est}} \to_p \sigma^2(\mathbf{w}^*).$$

Last, we show that data integration can improve efficiency. By combining these four parts, we directly obtain Theorem 4.12.

PART 1. Recall that the estimator $\widehat{\theta}(\mathbf{w})$ can be written as:

$$\widehat{\theta}(\mathbf{w}) = \sum_{k=0}^{K} w_k \widehat{\theta}_k = \widehat{\theta}_0 + \sum_{k=1}^{K} w_k \widehat{b}_k,$$

where $\widehat{b}_k = \widehat{\theta}_k - \widehat{\theta}_0$. Define the true weighted parameter $\theta^*(\mathbf{w}) = \sum_{k=0}^{K} w_k \theta_k = \theta_0 + \sum_{k=1}^{K} w_k b_k$. We first show that $\sqrt{N}(\theta^*(\mathbf{w}^*) - \theta_0) = o(1)$. To see this, consider two cases based on Assumption 4.10: First, for $b_k \ll N^{-1/2}$, it follows immediately that: $\sqrt{N} b_k = o(1)$. In addition, for $b_k \gg N^{-1/2}$, recall the oracle weight definition:

$$w_k^* = \frac{(\frac{\sigma_k^2}{n_k} + \lambda b_k^2)^{-1}}{(\frac{\sigma_0^2}{n_0})^{-1} + \sum_{j=1}^{K} (\frac{\sigma_j^2}{n_j} + \lambda b_j^2)^{-1}}.$$

Under Assumption 4.9, we obtain: $w_k^* = O(1/(\lambda b_k^2 N)) = o(1)$, which implies:

$$\sqrt{N} w_k^* b_k = O(\sqrt{N} \frac{1}{\lambda b_k^2 N} b_k) = O(\frac{1}{\lambda |b_k| \sqrt{N}}) = o(1). \tag{C.4}$$

Therefore, under Assumption 4.10, we have:

$$\sqrt{N}(\theta^*(\mathbf{w}) - \theta_0) = o(1).$$

Combining this with Lemma B.2, we conclude that:

$$\sqrt{N}(\widehat{\theta}(\mathbf{w}^*) - \theta_0) \to_d \mathcal{N}(0, \sigma^2(\mathbf{w}^*)).$$

This finishes PART 1.

PART 2. Next, we analyze: $\sqrt{N}\{\widehat{\theta}(\mathbf{w}^*) - \widehat{\theta}_{\text{est}}\}$. By definition, we have

$$\widehat{\theta}(\mathbf{w}^*) - \widehat{\theta}_{\text{est}} = \widehat{\theta}_0 + \sum_{k=1}^{K} w_k^*(\widehat{\theta}_k - \widehat{\theta}_0) - \Big\{\widehat{\theta}_0 + \sum_{k=1}^{K} \widehat{w}_k(\widehat{\theta}_k - \widehat{\theta}_0)\Big\} = \sum_{k=1}^{K}(w_k^* - \widehat{w}_k)(\widehat{\theta}_k - \widehat{\theta}_0).$$

Again, we consider two scenarios: When $b_k \ll N^{-1/2}$, it can be verified from equation (B.1) that:

$$\sqrt{N}(\widehat{\theta}_k - \widehat{\theta}_0) = \sqrt{N}(\widehat{b}_k - b_k) + \sqrt{N}b_k = O_p(1).$$

Using Lemma B.4 and Assumption 4.3, we thus have:

$$\sqrt{N}(w_k^* - \widehat{w}_k)(\widehat{\theta}_k - \widehat{\theta}_0) = o_p(1).$$

For the case $b_k \gg N^{-1/2}$, we similarly decompose:

$$\sqrt{N}(w_k^* - \widehat{w}_k)(\widehat{\theta}_k - \widehat{\theta}_0) = \sqrt{N}(w_k^* - \widehat{w}_k)(\widehat{\theta}_k - \theta_k) - \sqrt{N}(w_k^* - \widehat{w}_k)(\widehat{\theta}_0 - \theta_0) + \sqrt{N}(w_k^* - \widehat{w}_k)b_k.$$

By Lemmas B.2 and Lemma B.4, the first two terms are $o_p(1)$. Thus, it suffices to study $\sqrt{N}(w_k^* - \widehat{w}_k)b_k$. Note that

$$\sqrt{N}(w_k^* - \widehat{w}_k)b_k = \sqrt{N}w_k^* b_k - \sqrt{N}\widehat{w}_k(b_k - \widehat{b}_k) - \sqrt{N}\widehat{w}_k\widehat{b}_k.$$

First, from the analysis in PART 1, we already have $\sqrt{N}w_k^* b_k = o_p(1)$ by equation (C.4). Next, recall:

$$\widehat{w}_k = \frac{(\frac{\widehat{\sigma}_k^2}{n_k} + \lambda\widehat{b}_k^2)^{-1}}{\left(\frac{\widehat{\sigma}_0^2}{n_0}\right)^{-1} + \sum_{j=1}^{K}(\frac{\widehat{\sigma}_j^2}{n_j} + \lambda\widehat{b}_j^2)^{-1}}.$$

Using equation (B.1) again, it can be shown that when $b_k \gg N^{-1/2}$, $1/(\widehat{b}_k\sqrt{N}) = o_p(1)$ when $b_k \gg N^{-1/2}$. As a result, it could be shown that

$$\widehat{w}_k = O_p\Big(\frac{1}{\lambda\widehat{b}_k^2 N}\Big)$$

since $\widehat{\sigma}_k^2 - \sigma_k^2 = o_p(1)$. This yields

$$\sqrt{N}\widehat{w}_k\widehat{b}_k = O_p\Big(\frac{1}{\lambda\widehat{b}_k\sqrt{N}}\Big) = o_p(1).$$

Hence, under Assumption 4.10, we obtain:

$$\sqrt{N}\Big(\widehat{\theta}(\mathbf{w}^*) - \widehat{\theta}_{\text{est}}\Big) = o_p(1),$$

which completes PART 2.

PART 3. Subsequently, from Assumption 4.7, we have: $\widehat{\sigma}_k^2 - \sigma_k^2 = o_p(1)$. Combining this with Lemma B.4, we have: $\widehat{\sigma}_{\text{est}}^2 \to_p \sigma^2(\mathbf{w}^*)$, which completes PART 3.

PART 4. Recall that the loss function is defined as: $\mathcal{L}(\mathbf{w}) = \sum_{k=0}^{K} w_k^2 \frac{\sigma_k^2}{n_k} + \lambda \sum_{k=0}^{K} w_k^2 b_k^2$. The corresponding optimization problem is given by:

$$\mathbf{w}^* = \arg\min_{\mathbf{w}} \mathcal{L}(\mathbf{w}), \quad \text{subject to} \quad w_k \geq 0 \text{ for all } k, \quad \sum_{k=0}^{K} w_k = 1. \tag{C.5}$$

Clearly, $(1, 0, \ldots, 0)$ is a feasible solution to problem (C.5), with the corresponding loss equal to σ_0^2/n_0. As a consequence, to prove Theorem 4.12, it suffices to show that there exists a feasible solution $\mathbf{w}'$ satisfying $\mathcal{L}(\mathbf{w}') < \sigma_0^2/n_0$.

Suppose that the bias scale of the kth site satisfies $b_k \ll N^{-1/2}$. We define $\mathbf{w}' = (w_j')$ as follows:

$$w_0' = \frac{\sigma_k^2/n_k}{\sigma_0^2/n_0 + \sigma_k^2/n_k}, \quad w_k' = \frac{\sigma_0^2/n_0}{\sigma_0^2/n_0 + \sigma_k^2/n_k}, \quad w_j' = 0 \text{ for all } j \neq 0, k.$$

Clearly, $\mathbf{w}'$ is feasible, and it could be computed that:

$$\mathcal{L}(\mathbf{w}') = \frac{(\sigma_0^2/n_0)(\sigma_k^2/n_k)}{\sigma_0^2/n_0 + \sigma_k^2/n_k} + \lambda w_k'^2 b_k^2 = \frac{1}{\frac{1}{\sigma_0^2/n_0} + \frac{1}{\sigma_k^2/n_k}}(1 + o(1)).$$

This yields:

$$\frac{\mathcal{L}(\mathbf{w}')}{\sigma_0^2/n_0} = \frac{\sigma_k^2/n_k}{\sigma_0^2/n_0 + \sigma_k^2/n_k}(1 + o(1)) < 1.$$

This finishes the whole theorem proof.

C.4. Proof of Theorem 4.13.

Note that the conservative weight $\widetilde{w}_k$ does not necessarily converge to the oracle weight w_k^*. Denote $U_0 = 0$, $U_k = \mathcal{C}\sqrt{\sigma_k^2/n_k + \sigma_0^2/n_0}$ for $1 \leq k \leq K$, we introduce the auxiliary weights

$$\widetilde{w}_k^* = \frac{\left\{\frac{\sigma_k^2}{n_k} + \lambda\big(|b_k| + U_k\big)^2\right\}^{-1}}{\left(\frac{\sigma_0^2}{n_0}\right)^{-1} + \sum_{j=1}^K \left\{\frac{\sigma_j^2}{n_j} + \lambda\big(|b_j| + U_j\big)^2\right\}^{-1}} \text{ for } 0 \leq k \leq K.$$

Using these weights, we define the auxiliary estimator $\widehat{\theta}\big(\widetilde{\mathbf{w}}^*\big) = \sum_{k=0}^K \widetilde{w}_k^* \widehat{\theta}_k$. The proof has two steps. In the first step, we show that $\big|\widetilde{w}_k - \widetilde{w}_k^*\big| = o_p(1)$. In the second step, we establish Theorem 4.13 with the help of $\widetilde{w}_k^*$ and $\widehat{\theta}\big(\widetilde{\mathbf{w}}^*\big)$.

STEP 1. Recall the definition of $\widetilde{w}_k$, denote $\widehat{U}_0 = 0$, $\widehat{U}_k = \mathcal{C}\sqrt{\widehat{\sigma}_k^2/n_k + \widehat{\sigma}_0^2/n_0}$ for $1 \leq k \leq K$, we have

$$\widetilde{w}_k = \frac{\left\{\frac{\widehat{\sigma}_k^2}{n_k} + \lambda\big(|\widehat{b}_k| + \widehat{U}_k\big)^2\right\}^{-1}}{\left(\frac{\widehat{\sigma}_0^2}{n_0}\right)^{-1} + \sum_{j=1}^K \left\{\frac{\widehat{\sigma}_j^2}{n_j} + \lambda\big(|\widehat{b}_j| + \widehat{U}_j\big)^2\right\}^{-1}} \text{ for } 0 \leq k \leq K.$$

Then we have

$$\begin{aligned} \big|\widetilde{b}_k^2 - (|b_k| + U_k)^2\big| &= \big|(|\widehat{b}_k| + \widehat{U}_k)^2 - (|b_k| + U_k)^2\big| \\ &\leq \big\{(|\widehat{b}_k| + \widehat{U}_k) - (|b_k| + U_k)\big\}^2 + 2(|b_k| + U_k)\big|(|\widehat{b}_k| + \widehat{U}_k) - (|b_k| + U_k)\big|. \end{aligned} \tag{C.6}$$

It could be proved that

$$|\widehat{U}_k - U_k|^2 \leq \mathcal{C}^2\Big|\Big(\frac{\widehat{\sigma}_k^2}{n_k} + \frac{\widehat{\sigma}_0^2}{n_0}\Big) - \Big(\frac{\sigma_k^2}{n_k} + \frac{\sigma_0^2}{n_0}\Big)\Big| \leq \mathcal{C}^2\Big(\Big|\frac{\widehat{\sigma}_k^2}{n_k} - \frac{\sigma_k^2}{n_k}\Big| + \Big|\frac{\widehat{\sigma}_0^2}{n_0} - \frac{\sigma_0^2}{n_0}\Big|\Big).$$

This yields

$$|\widehat{U}_k - U_k| \leq \mathcal{C}\Big(\Big|\frac{\widehat{\sigma}_k^2}{n_k} - \frac{\sigma_k^2}{n_k}\Big| + \Big|\frac{\widehat{\sigma}_0^2}{n_0} - \frac{\sigma_0^2}{n_0}\Big|\Big)^{1/2}.$$

Because we assume $\mathcal{C} \to \infty$ and $\mathcal{C}\sqrt{|\widehat{\sigma}_k^2 - \sigma_k^2|} = \mathcal{O}_p(1)$ for every $k = 0, \ldots, K$, then we have

$$\Big|(|\widehat{b}_k| + \widehat{U}_k) - (|b_k| + U_k)\Big| \leq \Big||\widehat{b}_k| - |b_k|\Big| + \big|\widehat{U}_k - U_k\big| \leq |\widehat{b}_k - b_k| + |\widehat{U}_k - U_k| = \mathcal{O}_p\big(\frac{1}{\sqrt{N}}\big).$$

Substituting the bound above into equation (C.6) yields

$$\Big|(|\widehat{b}_k|+\widehat{U}_k)^2-(|b_k|+U_k)^2\Big| = \mathcal{O}_p\big(\frac{1}{N}\big)+\Big\{|b_k|+\mathcal{C}\Omega_p\big(\frac{1}{\sqrt{N}}\big)\Big\}\mathcal{O}_p\big(\frac{1}{\sqrt{N}}\big).$$

Next, by repeating the key steps in the proof of Lemma B.4, we can readily show that $\widetilde{w}_k-\widetilde{w}_k^*=o_p(1)$. The detailed proof is therefore omitted.

STEP 2. Note that $|b_k|+U_k \gg N^{-1/2}$, then it could be verified that

$$\widetilde{w}_k^* \lesssim \frac{\lambda^{-1}\Big\{(|b_k|+U_k)^2\Big\}^{-1}}{N+\sum_j\Big\{\lambda\big(|b_j|+U_j\big)^2\Big\}^{-1}} = \mathcal{O}\Big(\frac{1}{\lambda(|b_k|+U_k)^2N}\Big)=o(1),$$

and

$$\sqrt{N}\widetilde{w}_k^* b_k = \mathcal{O}\Big(\frac{b_k}{\lambda(|b_k|+U_k)^2\sqrt{N}}\Big)=o(1).$$

Similarly, we can prove that

$$\widetilde{w}_k = O_p\Big(\frac{1}{\lambda(|\widehat{b}_k|+\widehat{U}_k)^2N}\Big)$$

and

$$\sqrt{N}\widetilde{w}_k\widehat{b}_k = O_p\Big(\frac{\widehat{b}_k}{\lambda(|\widehat{b}_k|+\widehat{U}_k)^2\sqrt{N}}\Big)=o_p(1).$$

Next, by repeating the key steps in the proof of Theorem 4.12, it could be verified that

$$\sqrt{N}(\widehat{\theta}(\widetilde{\mathbf{w}}^*)-\theta_0)\to_d \mathcal{N}\big(0,\sigma^2(\widetilde{\mathbf{w}}^*)\big), \quad \sqrt{N}(\widehat{\theta}(\widetilde{\mathbf{w}}^*)-\widehat{\theta}_{\text{cons}})=o_p(1).$$

Finally, note that

$$\widehat{\sigma}^2_{\text{cons}} \to_p \sigma^2(\widetilde{\mathbf{w}}^*)$$

by Theorem 4.6 and using the fact that $\widetilde{w}_k-\widetilde{w}_k^*=o_p(1)$. By combining these three results, we finish the proof.